Journal of Machine Learning Research 10 (2009) 107-140 Submitted 2/08; Revised 9/08; Published 1/09

Refinement of Reproducing Kernels

Yuesheng Xu YXU06@SYR.EDU
Haizhang Zhang HZHANG12@SYR.EDU
Department of Mathematics
Syracuse University
Syracuse, NY 13244, USA

Editor: John Shawe-Taylor

Abstract

We continue our recent study on constructing a *refinement kernel* for a given kernel so that the reproducing kernel Hilbert space associated with the refinement kernel contains that with the original kernel as a subspace. To motivate this study, we first develop a refinement kernel method for learning, which gives an efficient algorithm for updating a learning predictor. Several characterizations of refinement kernels are then presented. It is shown that a nontrivial refinement kernel for a given kernel always exists if the input space has an infinite cardinal number. Refinement kernels for translation invariant kernels and Hilbert-Schmidt kernels are investigated. Various concrete examples are provided.

Keywords: reproducing kernels, reproducing kernel Hilbert spaces, learning with kernels, refinement kernels, translation invariant kernels, Hilbert-Schmidt kernels

1. Introduction

In our recent work (Xu and Zhang, 2007), we studied characterizations of a refinable kernel which offers a convenient way of enlarging its reproducing kernel Hilbert space (RKHS). Appropriately expanding a given RKHS is needed in learning theory when the given space is not adequate for a specific purpose. We will discuss this point in depth later. With a refinable kernel, a wavelet-like kernel or a *kernellet* was introduced in Xu and Zhang (2007). As pointed out there, the refinable kernel leaves out two important classes of kernels. Neither the Gaussian kernels nor kernels having finite dimensional feature spaces are refinable. Due to important applications of these classes of kernels, there is a need to develop a general method of enlarging a RKHS besides the particular one given by a refinable kernel. It is this need that leads to the study presented in this paper of *refinement* kernels for a given kernel.

We first review necessary notions related to kernels. Let X be a nonempty prescribed set called an input space. For $n \in \mathbb{N}$, we let $\mathbb{N}_n := \{1, 2, \ldots, n\}$. A *kernel* K on X is a function from $X \times X$ to the field $\mathbb{C}$ of complex numbers such that for any finite set of inputs $\mathbf{x} := \{x_j : j \in \mathbb{N}_n\} \subseteq X$ the matrix

$$K[\mathbf{x}] := [K(x_j, x_k) : j, k \in \mathbb{N}_n] \tag{1}$$

is hermitian and positive semi-definite. Kernels are important in learning theory as they are used to measure the similarity between inputs in X (Evgeniou et al., 2000; Schölkopf and Smola, 2002; Shawe-Taylor and Cristianini, 2004; Vapnik, 1998). A *reproducing kernel Hilbert space* (RKHS) on

X is a Hilbert space of functions on X for which point evaluations are continuous linear functionals (Aronszajn, 1950).

There is a bijective correspondence between the set of kernels on X and that of reproducing kernel Hilbert spaces (RKHS) on X. In particular, for each kernel K on X there is a unique RKHS $\mathcal{H}_K$ such that

$$K(\cdot,x)\in\mathcal{H}_K,\ \text{ for all } x\in X \tag{2}$$

and for all $f\in\mathcal{H}_K$ there holds

$$f(x)=(f,K(\cdot,x))_{\mathcal{H}_K},\ \ x\in X, \tag{3}$$

where $(\cdot,\cdot)_{\mathcal{H}_K}$ denotes the inner product on $\mathcal{H}_K$. Moreover, the linear span of $\{K(\cdot,x):x\in X\}$ is dense in $\mathcal{H}_K$, namely,

$$\mathcal{H}_K=\overline{\operatorname{span}}\{K(\cdot,x):x\in X\}, \tag{4}$$

and the inner product on $\mathcal{H}_K$ is determined by

$$(K(\cdot,y),K(\cdot,x))_{\mathcal{H}_K}=K(x,y),\ \ x,y\in X. \tag{5}$$

Conversely, for each RKHS $\mathcal{H}$ on X there exists exactly one kernel K on X such that (2) and (3) hold true with $\mathcal{H}_K$ replaced by $\mathcal{H}$. Equation (3) is interpreted as that a function in $\mathcal{H}_K$ can be *reproduced* through its inner product with the kernel K. For this reason, K is often called the *reproducing kernel* of $\mathcal{H}_K$.

The main purpose of this study is to investigate kernels K and G on X so that

$$\mathcal{H}_K\preceq\mathcal{H}_G \tag{6}$$

in the sense that $\mathcal{H}_K\subseteq\mathcal{H}_G$ and for all $f,g\in\mathcal{H}_K$, $(f,g)_{\mathcal{H}_K}=(f,g)_{\mathcal{H}_G}$. For an existing kernel K, we call a kernel G satisfying (6) a *refinement kernel* for K. If in addition, $\mathcal{H}_G$ contains $\mathcal{H}_K$ as a proper subspace, then we call G a *nontrivial* refinement kernel for K. The inclusion (6) was first considered by Aronszajn (1950). It was proved there that (6) holds true if and only if $L:=G-K$ remains a kernel on X and $\mathcal{H}_K\cap\mathcal{H}_L=\{0\}$.

Our interest in refinement kernels is motivated by the widely used regularized learning algorithm, which and its variations have attracted much attention in the literature (see, for example, Bousquet and Elisseeff, 2002; Cucker and Smale, 2002; Micchelli and Pontil, 2005a,b; Mukherjee et al., 2006; Schölkopf and Smola, 2002; Smale and Zhou, 2003; Steinwart and Scovel, 2005; Vapnik, 1998; Wahba, 1999; Walder et al., 2006; Ying and Zhou, 2007; Zhang, 2004, and the references cited therein). The algorithm aims at inferring from a finite set of training data $\mathbf{z}:=\{(x_j,y_j):j\in\mathbb{N}_m\}\subseteq X\times\mathbb{C}$ a function f_0 on X so that $f_0(x)$ would yield a meaningful output of an input $x\in X$. For a positive regularization parameter μ and the norm $\|\cdot\|_{\mathcal{H}_K}$ on $\mathcal{H}_K$, we set for each $f\in\mathcal{H}_K$

$$\mathcal{E}_{K,\mu}(f):=\sum_{j\in\mathbb{N}_m}|f(x_j)-y_j|^2+\mu\|f\|^2_{\mathcal{H}_K}.$$

The learning algorithm then outputs a predictor f_0 as the minimizer of an error functional:

$$f_0=\min_{f\in\mathcal{H}_K}\mathcal{E}_{K,\mu}(f). \tag{7}$$

The behavior of the predictor f_0 depends on the choice of the regularization parameter μ and as well as the RKHS. We will not consider the choice of μ in this paper. Rather, we will focus on the issue of expanding the RKHS associated with the original kernel K.

There are two possible situations where one may desire to find a nontrivial refinement kernel G for K in (7). The first happens when the predictor f_0 obtained from (7) does not work in a satisfactory way. One may hence be forced to replace K with a kernel G hoping that the corresponding learning algorithm would yield a better predictor. This is possible only if $\mathcal{H}_G$ is larger than $\mathcal{H}_K$. In other words, if the current RKHS underfits, then a refinement kernel may lead to a better predictor. The second situation occurs when the old training data $\mathbf{z}$ is expanded to be a new training data by adding to $\mathbf{z}$ more new samples from $X \times \mathbb{C}$. Since more information is available as the training data is increased, it is reasonable for people to expect a better predictor, which could be achieved by searching in a larger RKHS. This accounts for another reason one might want to find a refinement kernel for K.

In a recent paper (Xu and Zhang, 2007), we introduced a method of updating kernels via a composition of the kernel with a bijective mapping γ of the input space. Specifically, with a selected positive constant λ, we define for a kernel K on X a new kernel

$$G(x,y) := \lambda K(\gamma(x),\gamma(y)), \quad x,y \in X \tag{8}$$

and call K a γ-*refinable kernel* if (8) gives a nontrivial refinement kernel G for K. Various characterizations and many examples of refinable kernels were provided in Xu and Zhang (2007). The work was motivated by refinable functions in the context of wavelet analysis (Daubechies, 1992). As mentioned earlier, a purpose of the current study is to resolve two remaining questions in Xu and Zhang (2007). One is that a kernel is never refinable if it has a finite dimensional feature space. The other is that the commonly used Gaussian kernels are not refinable either. On the other hand, we know that kernels with a finite dimensional feature space, such as finite dot-product kernels (FitzGerald et al., 1995), and Gaussian kernels are important in learning (Micchelli and Pontil, 2005a; Schölkopf and Smola, 2002; Steinwart and Scovel, 2005; Walder et al., 2006). We would like to find nontrivial refinement kernels for them by considering general methods of updating kernels besides the particular one (8).

We organize this paper in six sections. Before delving into technical analysis of refinement kernels, we further motivate our study by proposing a refinement kernel method for learning in the next section. In Section 3, we present three basic characterizations of a refinement kernel. The first characterization is due to Aronszajn (1950), the second comes from a modification of a result in Xu and Zhang (2007) and the third result which is completely new serves as a base for further study in the remaining sections. We discuss in Section 4 the existence of a refinement kernel, and desired properties of kernels preserved by a refinement process. In particular, it will be shown that a nontrivial refinement kernel always exists if the input space contains infinite elements. In Sections 5 and 6, we study refinement kernels for translation invariant kernels and Hilbert-Schmidt kernels, respectively.

2. A Refinement Kernel Method for Learning

This section is devoted to development of learning algorithms based on refinement kernels. Suppose that a learning algorithm with kernel K has been given, that is, we have had a minimizer in the RKHS $\mathcal{H}_K$. But somehow we find that the minimizer is not good enough for a specific purpose. We then want to make a new search for a new minimizer in a larger RKHS $\mathcal{H}_G$, where G is a refinement kernel for K. We will demonstrate how a new search is done by making use of the previously computed results for the kernel K and the corresponding minimizer. We will refer to the methods described in this section as *refinement kernel methods* for learning.

For simplicity of presentation, we work only with real numbers in this section. Let K be a kernel on the input space X and $\mathbf{z} := \{(x_j, y_j) : j \in \mathbb{N}_m\} \subseteq X \times \mathbb{R}$ a finite set of sample data. We return to the learning algorithm described in the introduction which has the form

$$\min_{f \in \mathcal{H}_K} \left\{ \sum_{j \in \mathbb{N}_m} |f(x_j) - y_j|^2 + \mu \|f\|^2_{\mathcal{H}_K} \right\}. \tag{9}$$

The *representer theorem* (Kimeldorf and Wahba, 1971; Schölkopf et al., 2001; Schölkopf and Smola, 2002) in learning theory ensures that the minimizer $f_0 \in \mathcal{H}_K$ of (9) has the form

$$f_0 = \sum_{j \in \mathbb{N}_m} c_j K(\cdot, x_j).$$

In the above equation the vector $\mathbf{c} := [c_j : j \in \mathbb{N}_m]^T$ satisfies the linear system

$$(\mu I_m + K[\mathbf{x}])\mathbf{c} = \mathbf{y}, \tag{10}$$

where I_m denotes the $m \times m$ identity matrix, $\mathbf{x} := [x_j : j \in \mathbb{N}_m]^T$ and $\mathbf{y} := [y_j : j \in \mathbb{N}_m]^T$.

Suppose that the minimizer f_0 is not satisfactory and we need to have a new search in a larger RKHS. We assume that a refinement kernel G for K has been chosen and the training data $\mathbf{z}$ has been expanded to $\mathbf{z} \cup \mathbf{z}'$, where $\mathbf{z}' := \{(x'_k, y'_k) : k \in \mathbb{N}_q\} \subseteq X \times \mathbb{R}$. A new predictor can be obtained as the minimizer of

$$\min_{g \in \mathcal{H}_G} \left\{ \sum_{j \in \mathbb{N}_m} |g(x_j) - y_j|^2 + \sum_{k \in \mathbb{N}_q} |g(x'_k) - y'_k|^2 + \mu \|g\|^2_{\mathcal{H}_G} \right\}. \tag{11}$$

The purpose of this section is to develop an algorithm for efficiently solving (11) by using the existing information of the original minimization (9).

We proceed it in two steps.

2.1 Fixed Training Data

In this subsection, we assume that the training data set remains unchanged. Suppose that f_0 has been obtained, that is, linear system (10) has been solved, and one wishes to refine the kernel K in (9) to improve the predictor f_0. We consider a refinement kernel G for K for which the orthogonal complement of $\mathcal{H}_K$ in $\mathcal{H}_G$ is finite dimensional. We see from Aronszajn (1950) that there exist linearly independent functions ψ_j, $j \in \mathbb{N}_p$, on X, none of which lies in $\mathcal{H}_K$ such that the kernel $L := G - K$ has the form

$$L(x, y) := \sum_{j \in \mathbb{N}_p} \psi_j(x)\psi_j(y), \quad x, y \in X.$$

For instance, if K is the Gaussian kernel on $\mathbb{R}^d$ then L can be chosen as a finite dot-product kernel

$$\sum_{n \in \mathbb{Z}_+} a_n (x, y)^n, \quad x, y \in \mathbb{R}^d$$

or a *finite complex sinusoid kernel*

$$\sum_{n\in\mathbb{Z}^d} b_n e^{i(n,x-y)},\ \ x,y\in\mathbb{R}^d,$$

where a and b are a nonnegative function on $\mathbb{Z}_+ := \mathbb{N}\cup\{0\}$ and $\mathbb{Z}^d$, respectively, with finite supports.

Using the refinement kernel G, we shall obtain a new predictor g_0 in a larger RKHS $\mathcal{H}_G$ that is the minimizer of

$$\min_{g\in\mathcal{H}_G}\left\{\sum_{j\in\mathbb{N}_m}|g(x_j)-y_j|^2+\mu\|g\|^2_{\mathcal{H}_G}\right\}.$$

By the representer theorem, the predictor g_0 is of the form

$$g_0=\sum_{j\in\mathbb{N}_m} d_j G(\cdot,x_j),$$

where $\mathbf{d} := [d_j : j\in\mathbb{N}_m]^T$ satisfies

$$(\mu I_m+K[\mathbf{x}]+L[\mathbf{x}])\mathbf{d}=\mathbf{y}. \tag{12}$$

Suppose that the computational results in solving (10) have been stored. Since in general $K[\mathbf{x}]$ is a dense positive semi-definite matrix, linear system (10) is usually solved by the Cholesky factorization method (Golub and van Loan, 1996). The method works at a cost of $O(m^3)$ multiplications of real numbers. We hence assume that we have obtained the Cholesky factorization of $\mu I_m+K[\mathbf{x}]$. The decomposition (12) enables us to solve a linear system whose coefficient matrix is $\mu I_m+K[\mathbf{x}]$ using only $O(m^2)$ multiplications.

Now we have to solve the new linear system (12). Instead of spending another $O(m^3)$ multiplications, we wish to make use of the stored computational results from solving (10) to reduce the computational complexity in solving (12). Specifically, for each $j\in\mathbb{N}_p$, we let $\psi_j(\mathbf{x}) := [\psi_j(x_k) : k\in\mathbb{N}_m]^T$ and let $\mathbf{e}_j$ denote the vector satisfying

$$(\mu I_m+K[\mathbf{x}])\mathbf{e}_j=\psi_j(\mathbf{x}). \tag{13}$$

We also need the vector

$$\beta := [-\psi_j(\mathbf{x})^T\mathbf{c} : j\in\mathbb{N}_p] \tag{14}$$

and the $p\times p$ matrix B defined by

$$B_{jk} := \psi_j(\mathbf{x})^T\mathbf{e}_k,\ \ j,k\in\mathbb{N}_p. \tag{15}$$

Since by (13) there holds for each $j,k\in\mathbb{N}_p$ that $B_{jk}=\mathbf{e}_j^T(\mu I_m+K[\mathbf{x}])\mathbf{e}_k$ and $\mu I_m+K[\mathbf{x}]$ is symmetric and strictly positive definite, B is symmetric and positive semi-definite. As a consequence, I_p+B is invertible and we are allowed to introduce another vector $\alpha := [\alpha_j : j\in\mathbb{N}_p]^T\in\mathbb{R}^p$ as the unique solution of

$$(I_p+B)\alpha=\beta. \tag{16}$$

Proposition 1 *If vectors* $\mathbf{e}_j\in\mathbb{R}^m$, $j\in\mathbb{N}_p$ *are defined by (13),* $\mathbf{c}\in\mathbb{R}^m$ *by (10) and* $\alpha\in\mathbb{R}^p$ *by (16), then the solution* $\mathbf{d}$ *of linear system (12) is given by*

$$\mathbf{d}=\mathbf{c}+\sum_{j\in\mathbb{N}_p}\alpha_j\mathbf{e}_j. \tag{17}$$

If $\mathcal{N}(\mathbf{d})$ *denotes the number of multiplications required for computing* $\mathbf{d}$*, then*

$$\mathcal{N}(\mathbf{d}) = O(pm^2 + p^2m + p^3).$$

Proof Let $\mathbf{d}' := \mathbf{d} - \mathbf{c}$. It is clear by (10) and (12) that $\mathbf{d}$ satisfies (12) if and only if

$$(\mu I_m + K[\mathbf{x}])\mathbf{d}' + L[\mathbf{x}]\mathbf{d} = 0. \tag{18}$$

To prove the first statement of this proposition, it suffices to show that the vector $\mathbf{d}$ defined by (17) satisfies Equation (18). We denote by δ the left-hand side of (18). Substituting (17) into the left-hand side of (18) and noting that $L[\mathbf{x}] = \sum_{j\in\mathbb{N}_p} \psi_j(\mathbf{x})\psi_j(\mathbf{x})^T$, we obtain that

$$\delta = \sum_{j\in\mathbb{N}_p} \alpha_j\psi_j(\mathbf{x}) + \sum_{j\in\mathbb{N}_p} \psi_j(\mathbf{x})[\psi_j(\mathbf{x})^T\mathbf{c} + \sum_{k\in\mathbb{N}_p} \alpha_k\psi_j(\mathbf{x})^T\mathbf{e}_k].$$

By using (14) and (15), we have that

$$\delta = \sum_{j\in\mathbb{N}_p} \psi_j(\mathbf{x})\left[\alpha_j - \beta_j + \sum_{k\in\mathbb{N}_p} B_{jk}\alpha_k\right]. \tag{19}$$

Noting that α satisfies linear system (16) we observe for all $j \in \mathbb{N}_p$ that

$$\alpha_j - \beta_j + \sum_{k\in\mathbb{N}_p} B_{jk}\alpha_k = 0.$$

Combining this equation with (19) yields that $\delta = 0$. That is, the vector $\mathbf{d}$ defined by (17) satisfies Equation (18).

To prove the second statement, we make use of the assumption that the result of the Cholesky factorization of $\mu I_m + K[\mathbf{x}]$ has been computed and stored and we enumerate the additional number of multiplications required for computing the solution $\mathbf{d}$. Solving p linear systems (13) needs $O(pm^2)$ number of multiplications. Computing vector β and matrix B requires pm and $\frac{p(p+1)}{2}m$ multiplications respectively. Solving (16) costs $O(p^3)$ multiplications and finally, computing $\mathbf{d}$ by (17) requires pm multiplications. Summing these costs together yields the number of multiplications required to solve (12). ■

We remark that in applications, the number m of sample data is much larger than the number p of the dimension of the difference space $\mathcal{H}_L$. Therefore, under the condition $p \ll m$, we know from Proposition 1 that computing the solution $\mathbf{d}$ of the linear system (12) requires $O(m^2)$ additional number of multiplications. This is a big saving in comparison to $O(m^3)$ number of multiplications if the linear system (12) is solved directly by the Cholesky factorization without using the refinement kernel method. In other words, the use of the refinement kernel method reduces the number of multiplications from $O(m^3)$ to $O(m^2)$.

Moreover, we observe that most of the computational costs are used for solving (13), which is clearly independent of the output $\mathbf{y}$. Therefore, if $\mathbf{x}$ remains fixed for different applications, which is the case in many practical scenarios such as image analysis and processing of signals of the same size, then (13) can be calculated in advance and stored for repeated use. Taking this advantage, we only need $O(m)$ additional number of multiplications to solve (12) in order to obtain an updated predictor.

2.2 Expanded Training Data

We assume in this subsection that the training data $\mathbf{z}$ has been expanded to $\mathbf{z}\cup\mathbf{z}'$ while the kernel K remains the same. A new predictor $f_1\in\mathcal{H}_K$ is obtained by solving the minimization problem

$$\min_{f\in\mathcal{H}_K}\sum_{j\in\mathbb{N}_m}|f(x_j)-y_j|^2+\sum_{k\in\mathbb{N}_q}|f(x'_k)-y'_k|^2+\mu\|f\|^2_{\mathcal{H}_K}. \tag{20}$$

The predictor f_1 can be written in terms of the kernel K. To this end, we introduce two vectors

$$\mathbf{x}':=[x'_k:k\in\mathbb{N}_q]^T,\ \ \mathbf{y}':=[y'_k:k\in\mathbb{N}_q]^T,$$

and matrices

$$K[\mathbf{x},\mathbf{x}']:=[K(x_j,x'_k):j\in\mathbb{N}_m,k\in\mathbb{N}_q],\ \ K[\mathbf{x}',\mathbf{x}]:=K[\mathbf{x},\mathbf{x}']^T.$$

For notational simplicity, we also set $A:=\mu I_m+K[\mathbf{x}]$, $B:=\mu I_q+K[\mathbf{x}']$ and $C:=K[\mathbf{x},\mathbf{x}']$. By the representer theorem, the minimizer f_1 of (20) is given by

$$f_1=\sum_{j\in\mathbb{N}_m}d_jK(\cdot,x_j)+\sum_{k\in\mathbb{N}_q}d'_kK(\cdot,x'_k),$$

where $\mathbf{d}:=[d_j:j\in\mathbb{N}_m]^T$ and $\mathbf{d}':=[d'_k:k\in\mathbb{N}_q]^T$ satisfy

$$\begin{bmatrix} A & C\\ C^T & B\end{bmatrix}\begin{bmatrix}\mathbf{d}\\ \mathbf{d}'\end{bmatrix}=\begin{bmatrix}\mathbf{y}\\ \mathbf{y}'\end{bmatrix}. \tag{21}$$

The above linear system generally cost $O((m+q)^3)$ number of multiplications to solve by using the Cholesky factorization. With the known Cholesky factorization of $\mu I_m+K[\mathbf{x}]$, we propose a method to solve system (21) with reduction in the computational costs.

To solve system (21), we first find the $m\times q$ matrix M that satisfies

$$AM=C. \tag{22}$$

Note that

$$\begin{bmatrix} A & 0\\ 0 & B-C^TM\end{bmatrix}=\begin{bmatrix} I_m & 0\\ -C^TA^{-1} & I_q\end{bmatrix}\begin{bmatrix} A & C\\ C^T & B\end{bmatrix}\begin{bmatrix} I_m & -A^{-1}C\\ 0 & I_q\end{bmatrix}.$$

Thus $B-C^TM$ is symmetric and strictly positive definite. Consequently, we can solve the following system for a unique vector $\eta\in\mathbb{R}^q$

$$(B-C^TM)\eta=C^T\mathbf{c}-\mathbf{y}' \tag{23}$$

by using again the Cholesky factorization.

Proposition 2 *The solution* $\mathbf{d}$, $\mathbf{d}'$ *for linear system (21) is given as*

$$\mathbf{d}:=\mathbf{c}+M\eta,\ \mathbf{d}':=-\eta. \tag{24}$$

Moreover, if $\mathcal{N}(\mathbf{d},\mathbf{d}')$ *denotes the number of multiplications required for computing both* $\mathbf{d}$ *and* $\mathbf{d}'$*, then*

$$\mathcal{N}(\mathbf{d},\mathbf{d}')=O(qm^2+q^2m+q^3).$$

Proof The first statement of this theorem follows by a direct computation using (10), (21), (22) and (23).

We now count the number of multiplications used for computing both $\mathbf{d}$ and $\mathbf{d}'$. Computing matrix M by solving the matrix Equation (22) needs $O(qm^2)$ number of multiplications. Finding η from the system (23) takes up $O(q^2m+q^3)$ number of multiplications. Computing $\mathbf{d}$ and $\mathbf{d}'$ by using (24) costs $O(qm)$ number of multiplications. Summing all these costs up proves the second part of this proposition. ■

We remark that under the assumption that $q \ll m$, we only need additional $O(m^2)$ number of multiplications to solve system (21) based on the known Cholesky factorization of matrix A from solving (10). Since (22) is independent of outputs, making use of this feature, if the inputs $\mathbf{x}$, $\mathbf{x}'$ remain unchanged in different applications, the computational costs can be further reduced as we have mentioned at the end of the last subsection.

2.3 The General Case

We now return to the general case where we have both a refinement kernel G for K such that $\mathcal{H}_{G-K}$ is p-dimensional, and an expanded training data $\mathbf{z}\cup\mathbf{z}'$. To make use of the computational result in solving (10) to compute the minimizer of (11), we divide the updating process into two steps. In step one, we fix the training data to be $\mathbf{z}$ and refine the kernel K to G. We then solve the system using the method described in Section 2.1. In step two, we fix the kernel to be G and expand $\mathbf{z}$ to $\mathbf{z}\cup\mathbf{z}'$ and solve the problem described in Section 2.2 with K replaced by G. Under reasonable hypotheses, the number of multiplications is $O(m^2)$. We state this result in the next proposition.

Proposition 3 *If $p \ll m$ and $q \ll m$ then the number of multiplications required for computing the minimizer of (11) using the algorithm described above is given by $O(m^2)$.*

The refinement kernel learning method allows us to reduce the number of multiplications from $O(m^3)$ to $O(m^2)$ by using the known Cholesky factorization of the matrix corresponding to the old kernel K.

Here we focus on the computational complexity of the refinement kernel method for the regularized learning algorithm in order to motivate the study of the refinement kernel. Although convergence and consistency of the refinement kernel method, and extensions of the method to other learning algorithms such as support vector machines are important, they are not the focus of this paper. They will be addressed in different occasions. The rest of this paper will be devoted to theoretical analysis of refinement kernels such as characterizations, existence and constructions.

3. Characterizations of Refinement Kernels

We present in this section several characterizations of refinement kernels. We begin with a review of a well-known characterization from Aronszajn (1950).

Lemma 4 *Let K,G be kernels on X. Then G is a refinement kernel for K if and only if $L := G-K$ is a kernel on X and $\mathcal{H}_K \cap \mathcal{H}_L = \{0\}$. If G is a refinement kernel for K then $\mathcal{H}_L$ is the orthogonal complement of $\mathcal{H}_K$ in $\mathcal{H}_G$, and G is a nontrivial refinement kernel for K if and only if L is not the zero kernel.*

Proof This is a direct consequence of Property 7 on page 345 and the theorem on page 353 of Aronszajn (1950). ■

In general, the conditions in the characterization presented in Lemma 4 are not easy to verify. Aiming at a characterization convenient for use, we next characterize refinement kernels in terms of feature maps for kernels, since most kernels are identified with their feature maps. A *feature map* for a kernel K on X is a mapping Φ from X to a Hilbert space $\mathcal{W}$ over $\mathbb{C}$ such that

$$K(x,y) = (\Phi(x), \Phi(y))_{\mathcal{W}}, \ \ x,y \in X. \tag{25}$$

The Hilbert space $\mathcal{W}$ is called a *feature space* for K. It is well-known that K is a kernel on X if and only if it can be represented as (25) for some mapping $\Phi : X \to \mathcal{W}$ (Schölkopf and Smola, 2002). We denote by $\Phi(X)$ the image of X under the mapping Φ, by $\overline{\text{span}}\,\Phi(X)$ the closure of the linear span of $\Phi(X)$ in $\mathcal{W}$, and by P_Φ the orthogonal projection from $\mathcal{W}$ to $\overline{\text{span}}\,\Phi(X)$. There is a well-known characterization (Micchelli and Pontil, 2005a; Opfer, 2006; Schölkopf and Smola, 2002; Shawe-Taylor and Cristianini, 2004; Xu and Zhang, 2007) of the RKHS $\mathcal{H}_K$ of K in terms of its feature maps.

Lemma 5 *If K is a kernel on X represented as (25) by a feature map Φ from X to $\mathcal{W}$, then $\mathcal{H}_K = \{(\Phi(\cdot), u)_{\mathcal{W}} : u \in \mathcal{W}\}$ with the inner product*

$$((\Phi(\cdot), u)_{\mathcal{W}}, (\Phi(\cdot), v)_{\mathcal{W}})_{\mathcal{H}_K} = (P_\Phi v, P_\Phi u)_{\mathcal{W}}, \ \ u,v \in \mathcal{W}. \tag{26}$$

We shall always assume in the application of Lemma 5 that

$$\overline{\text{span}}\,\Phi(X) = \mathcal{W} \tag{27}$$

since (25) remains valid if we replace $\mathcal{W}$ there with $\overline{\text{span}}\,\Phi(X)$. Convenience which results from this assumption is that P_Φ in (26) would become the identity operator on $\mathcal{W}$.

We next state a characterization of refinement kernels in terms of their feature maps. Recall that we call a linear operator T from Hilbert space $\mathcal{W}_1$ to Hilbert space $\mathcal{W}_2$ *isometric* if for each $u \in \mathcal{W}_1$, $\|Tu\|_{\mathcal{W}_2} = \|u\|_{\mathcal{W}_1}$. A linear operator T from Hilbert space $\mathcal{W}_1$ to Hilbert space $\mathcal{W}_2$ is called an *isomorphism* if it is a bijective isometric linear mapping from $\mathcal{W}_1$ to $\mathcal{W}_2$. If there is an isomorphism from $\mathcal{W}_1$ to $\mathcal{W}_2$, we say that $\mathcal{W}_1$ is *isomorphic* to $\mathcal{W}_2$.

Theorem 6 *Suppose that K is a kernel on X with a feature map $\Phi : X \to \mathcal{W}$ satisfying (27) and G is a kernel on X with a feature map $\Phi' : X \to \mathcal{W}'$ that satisfies*

$$\overline{\text{span}}\,\Phi'(X) = \mathcal{W}'.$$

Then G is a refinement kernel for K if and only if there exists a bounded linear operator $T : \mathcal{W}' \to \mathcal{W}$ such that

$$T\Phi'(x) = \Phi(x), \ \ x \in X \tag{28}$$

and the adjoint operator $T^* : \mathcal{W} \to \mathcal{W}'$ of T is isometric. Moreover, G is a nontrivial refinement kernel for K if and only if T in (28) is not injective.

Proof The proof for the case when $\mathcal{W}' = \mathcal{W}$ and G is given by (8) can be found in Xu and Zhang (2007) (see Theorems 6, 7 therein). Based on Lemma 5, the arguments used there can be extended in a direct way to prove the general result described here. ■

We now explain how the general context of Theorem 6 allows us to refine kernels with a finite dimensional feature space while the refinement approach (8) does not. Suppose that K has the feature map representation (25) and G is given by (8). Then it can be seen that G has the form

$$G(x,y) = (\lambda^{1/2}\Phi(\gamma(x)), \lambda^{1/2}\Phi(\gamma(y)))_{\mathcal{W}}, \ \ x,y \in X$$

with the feature space $\mathcal{W}$. By Theorem 6, G is a nontrivial refinement kernel for K if and only if there exists a bounded linear operator $T : \mathcal{W} \to \mathcal{W}$ such that

$$\lambda^{1/2} T\Phi(\gamma(x)) = \Phi(x), \ \ x \in X,$$

its adjoint T^* is isometric from $\mathcal{W}$ to $\mathcal{W}$, and T is not injective. By the latter two conditions, $\mathcal{W}$ must be isomorphic to a proper subspace of itself. This is impossible if $\mathcal{W}$ is finite dimensional. Therefore, we can not get a nontrivial refinement kernel by (8) if $\mathcal{W}$ is of finite dimension. On the other hand, by considering a general refinement process

$$G(x,y) = (\Phi'(x), \Phi'(y))_{\mathcal{W}'}, \ \ x,y \in X,$$

we have the freedom to choose $\mathcal{W}'$ different from $\mathcal{W}$. The fact that $\mathcal{W}$ is finite dimensional actually makes it easier to find $\mathcal{W}'$ such that $\mathcal{W}$ is isomorphic to a proper subspace of $\mathcal{W}'$. Examples of nontrivial refinement kernels for kernels with a finite dimensional feature space will be provided in Section 6.

Our next task is to present a characterization of refinement kernels in terms of their feature spaces defined by finite positive Borel measures. This result is crucial for our discussion later on translation invariant kernels and Hilbert-Schmidt kernels. Suppose that Y is a topological space and denote by $\mathcal{B}(Y)$ the set of finite positive Borel measures on Y. For each $\rho \in \mathcal{B}(Y)$ and $p \in [1,+\infty)$ we let $L^p(Y,\rho)$ denote the space of Borel measurable functions f on Y such that

$$\int_Y |f(\xi)|^p d\rho(\xi) < +\infty.$$

In particular, $L^2(Y,\rho)$ is a Hilbert space with the inner product

$$(f,g)_{L^2(Y,\rho)} := \int_Y f(\xi)\overline{g(\xi)} d\rho(\xi), \ \ f,g \in L^2(Y,\rho).$$

For two measures $\rho_1, \rho_2 \in \mathcal{B}(Y)$, ρ_1 is said to be *absolutely continuous* with respect to ρ_2, denoted as $\rho_1 \ll \rho_2$, if for each Borel subset $V \subseteq Y$ with $\rho_2(V) = 0$, we have $\rho_1(V) = 0$. By the Radon-Nikodym theorem (see, for example, Rudin, 1987, page 121), if $\rho_1 \ll \rho_2$ then there exists a nonnegative $h \in L^1(Y,\rho_2)$ such that there holds for each Borel subset $V \subseteq Y$

$$\rho_1(V) = \int_Y h(\xi)\chi_V(\xi) d\rho_2(\xi),$$

where χ_V denotes the characteristic function of V. We sometimes write the function h satisfying the above equation as $d\rho_1/d\rho_2$.

For $\mu,\nu \in \mathcal{B}(Y)$, we let

$$\omega := \frac{\mu+\nu}{2} + \frac{|\mu-\nu|}{2},$$

where $|\mu-\nu|$ denotes the *total variation measure* of $\mu-\nu$. For the definition and properties of total variations of signed measures, see Rudin (1987, page 116). We remark that $|\mu-\nu| \in \mathcal{B}(Y)$ and $\mu(V),\nu(V) \leq \omega(V)$ for all Borel subsets V of Y. It follows that μ,ν are absolutely continuous with respect to ω. We assume that a function $\phi : X \times Y \to \mathbb{C}$ has the property that for each $x \in X$, $\phi(x,\cdot) \in L^2(Y,\omega)$ and

$$\overline{\operatorname{span}}\{\phi(x,\cdot) : x \in X\} = L^2(Y,\omega). \tag{29}$$

Lemma 7 *If $\phi : X \times Y \to \mathbb{C}$ satisfies $\phi(x,\cdot) \in L^2(Y,\omega)$ for each $x \in X$ and condition (29), then $\phi(x,\cdot)$ lies in $L^2(Y,\mu)$ and $L^2(Y,\nu)$ for all $x \in X$, and* $\operatorname{span}\{\phi(x,\cdot) : x \in X\}$ *is dense in $L^2(Y,\mu)$ and $L^2(Y,\nu)$.*

Proof We present only the case for μ since the case for ν can be similarly handled. Since $\mu \ll \omega$, we may introduce a function $h_\mu := d\mu/d\omega$. By the fact that $\mu(V) \leq \omega(V)$ for each Borel $V \subseteq Y$, h_μ is less than or equal to 1 almost everywhere on Y with respect to ω. For $x \in X$, the assumption that $\phi(x,\cdot) \in L^2(Y,\omega)$ implies that it is Borel measurable. We also verify that

$$\int_Y |\phi(x,\xi)|^2 d\mu(\xi) = \int_Y |\phi(x,\xi)|^2 h_u(\xi) d\omega(\xi) \leq \int_Y |\phi(x,\xi)|^2 d\omega(\xi) < +\infty.$$

This yields that $\phi(x,\cdot) \in L^2(Y,\mu)$ for all $x \in X$.

Now we assume that $f \in L^2(Y,\mu)$ is orthogonal to $\phi(x,\cdot)$ for each $x \in X$, that is,

$$\int_Y \phi(x,\xi)\overline{f(\xi)} d\mu(\xi) = 0, \;\; x \in X.$$

In the above equation, we substitute $d\mu(\xi) = h_\mu(\xi) d\omega(\xi)$ to obtain that

$$\int_Y \phi(x,\xi)\overline{f(\xi)} h_\mu(\xi) d\omega(\xi) = 0, \;\; x \in X. \tag{30}$$

The function fh_μ belongs to $L^2(Y,\omega)$ since h_μ is less than or equal to 1 almost everywhere on Y with respect to ω. Thus, by condition (29), Equation (30) implies that $fh_\mu = 0$ with respect to ω. We then observe for each Borel subset $V \subseteq Y$ that

$$\int_V |f(\xi)| d\mu(\xi) = \int_V |f(\xi)| h_\mu(\xi) d\omega(\xi) = 0.$$

This ensures that f vanishes almost everywhere on Y with respect to μ. We conclude that $\operatorname{span}\{\phi(x,\cdot) : x \in X\}$ is dense in $L^2(Y,\mu)$. ■

By virtue of the above lemma, we introduce two kernels K_μ, K_ν by setting for all $x,y \in X$

$$K_\mu(x,y) := (\phi(x,\cdot),\phi(y,\cdot))_{L^2(Y,\mu)}, \;\; K_\nu(x,y) := (\phi(x,\cdot),\phi(y,\cdot))_{L^2(Y,\nu)}. \tag{31}$$

By Lemmas 5 and 7, functions in $\mathcal{H}_\mu := \mathcal{H}_{K_\mu}$ have the form

$$f_{\phi,\mu}(x) := (\phi(x,\cdot), f)_{L^2(Y,\mu)}, \;\; x \in X, \;\; f \in L^2(Y,\mu)$$

and the inner product on $\mathcal{H}_\mu$ is given by

$$(f_{\phi,\mu}, g_{\phi,\mu})_{\mathcal{H}_\mu} = (g, f)_{L^2(Y,\mu)}, \quad f, g \in L^2(Y,\mu).$$

Similar results hold for $\mathcal{H}_\nu := \mathcal{H}_{K_\nu}$.

We shall characterize the relation $\mathcal{H}_\mu \preceq \mathcal{H}_\nu$ in terms of a relation of the measures μ, ν. To this end, we write $\mu \preceq \nu$ to indicate that $\mu \ll \nu$, and $d\mu/d\nu$ equals 0 or 1 almost everywhere with respect to ν, that is,

$$\nu\left(Y \setminus \left\{x \in Y : \frac{d\mu}{d\nu}(x) = 0 \text{ or } 1\right\}\right) = 0.$$

Note that $\mu \preceq \nu$ if and only if there exists a Borel subset $E \subseteq Y$ such that $\mu(Y \setminus E) = 0$ and for each Borel subset $V \subseteq E$, $\mu(V) = \nu(V)$.

Theorem 8 *Suppose that $\phi : X \times Y \to \mathbb{C}$ satisfies (29) and K_μ, K_ν are defined by (31). Then $\mathcal{H}_\mu \preceq \mathcal{H}_\nu$ if and only if $\mu \preceq \nu$. If $\mu \preceq \nu$ then K_ν is a nontrivial refinement kernel for K_μ if and only if*

$$\nu(Y) - \mu(Y) > 0.$$

Proof Suppose that $\mathcal{H}_\mu \preceq \mathcal{H}_\nu$. Hence, by Lemma 5, for each $f \in L^2(Y,\mu)$ there exists some $g \in L^2(Y,\nu)$ such that

$$\int_Y \phi(x,\xi)\overline{f(\xi)}d\mu(\xi) = \int_Y \phi(x,\xi)\overline{g(\xi)}d\nu(\xi) \tag{32}$$

and

$$\int_Y |f(\xi)|^2 d\mu(\xi) = \int_Y |g(\xi)|^2 d\nu(\xi). \tag{33}$$

With the derivatives $h_u := d\mu/d\omega$ and $h_\nu := d\nu/d\omega$, Equation (32) is rewritten as

$$\int_Y \phi(x,\xi)\overline{f(\xi)}h_\mu(\xi)d\omega(\xi) = \int_Y \phi(x,\xi)\overline{g(\xi)}h_\nu(\xi)d\omega(\xi).$$

This together with the density condition (29) implies that $fh_\mu = gh_\nu$ almost everywhere on Y with respect to ω. Thus for each $f \in L^2(Y,\mu)$ there exists some $g \in L^2(Y,\nu)$ satisfying for all Borel $V \subseteq Y$ that

$$\int_V f(\xi)d\mu(\xi) = \int_V f(\xi)h_\mu(\xi)d\omega(\xi) = \int_V g(\xi)h_\nu(\xi)d\omega(\xi) = \int_V g(\xi)d\nu(\xi). \tag{34}$$

We claim that $\mu \ll \nu$. We assume to the contrary that there exists a Borel set $V \subseteq Y$ for which $\nu(V) = 0$ and $\mu(V) > 0$. Letting $f = \chi_V$ in (34) yields $\mu(V) = 0$, a contradiction. Set $h := d\mu/d\nu$. By (34), the function g satisfying (32) and (33) can be taken as $g := fh$. With this choice, we obtain from (33) for each $f \in L^2(Y,\mu)$ that

$$\int_Y |f(\xi)|^2 h(\xi)d\nu(\xi) = \int_Y |f(\xi)|^2 h^2(\xi)d\nu(\xi).$$

The above equation implies that h equals 1 or 0 almost everywhere on Y with respect to ν. Consequently, $\mu \preceq \nu$.

Conversely, we suppose that $\mu \preceq \nu$ and proceed the proof by using Theorem 6. To this end, we set $E := \{x \in Y : \frac{d\mu}{d\nu}(x) = 1\}$ and introduce a linear operator $T : L^2(Y,\nu) \to L^2(Y,\mu)$ by

$$Tf := f\chi_E, \quad f \in L^2(Y,\nu).$$

By the hypothesis $\mu \preceq \nu$, we have that $\mu(Y \setminus E) = 0$. This guarantees that for each $x \in X$, $T\phi(x,\cdot) = \phi(x,\cdot)$ in $L^2(Y,\mu)$. Note that for $g \in L^2(Y,\mu)$ and $f \in L^2(Y,\nu)$,

$$\begin{aligned}\int_Y g(\xi)\overline{(Tf)(\xi)}d\mu(\xi) &= \int_Y g(\xi)\overline{f(\xi)\chi_E(\xi)}d\mu(\xi) = \int_E g(\xi)\overline{f(\xi)}d\mu(\xi)\\ &= \int_E g(\xi)\overline{f(\xi)}d\nu(\xi) = \int_Y g(\xi)\chi_E(\xi)\overline{f(\xi)}d\nu(\xi).\end{aligned}$$

This ensures that the adjoint $T^* : L^2(Y,\mu) \to L^2(Y,\nu)$ of T is given by $T^*g = g\chi_E$, for $g \in L^2(Y,\mu)$. Moreover, it can be verified that

$$\int_Y |(T^*g)(\xi)|^2 d\nu(\xi) = \int_Y |g(\xi)\chi_E(\xi)|^2 d\nu(\xi) = \int_E |g(\xi)|^2 d\nu(\xi).$$

Since $d\mu/d\nu = 1$ on E and $\mu(Y \setminus E) = 0$, we get that

$$\int_E |g(\xi)|^2 d\nu(\xi) = \int_E |g(\xi)|^2 \frac{d\mu}{d\nu}(\xi) d\nu(\xi) = \int_E |g(\xi)|^2 d\mu(\xi) = \int_Y |g(\xi)|^2 d\mu(\xi).$$

Combining the above two equations yields that T^* is isometric. By Theorem 6, K_ν is a refinement kernel for K_μ.

If $\mu \preceq \nu$ then K_ν is a nontrivial refinement kernel for K_μ if and only if the operator T is not injective, that is, there exists $f \in L^2(Y,\nu)$ such that

$$\|f\|_{L^2(Y,\nu)} > 0 \ \text{ but } \ \|Tf\|_{L^2(Y,\mu)} = 0.$$

This is equivalent to that

$$\int_Y |f(\xi)|^2 d\nu(\xi) > 0 \ \text{ but } \ \int_E |f(\xi)|^2 d\nu(\xi) = 0.$$

Clearly, such an f exists if and only if $\nu(Y \setminus E) > 0$. Because

$$\nu(Y) - \mu(Y) = \nu(Y) - \mu(E) = \nu(Y) - \nu(E) = \nu(Y \setminus E),$$

we conclude the second statement of the theorem. ■

4. Existence of Refinement Kernels

With characterizations in Lemma 4 and Theorem 6, we shall consider existence of nontrivial refinement kernels and properties of kernels preserved by the refinement process. We let $\mathbb{C}^X$ denote the space of all the complex-valued functions on X.

Lemma 9 *A kernel K on X does not have a nontrivial refinement kernel if and only if $\mathcal{H}_K = \mathbb{C}^X$.*

Proof If $\mathcal{H}_K = \mathbb{C}^X$ then since for all kernels G on X, $\mathcal{H}_G \subseteq \mathbb{C}^X$, K does not have a nontrivial refinement kernel. Conversely, if $\mathcal{H}_K \neq \mathbb{C}^X$ then we choose an arbitrary function $\varphi \in \mathbb{C}^X \setminus \mathcal{H}_K$ and define the kernel

$$G(x,y) := K(x,y) + \varphi(x)\overline{\varphi(y)}, \ \ x,y \in X.$$

It is clear that $L := G - K$ is a kernel on X since it has a feature map $\varphi : X \to \mathbb{C}$. Moreover, $\mathcal{H}_L = \operatorname{span}\{\varphi\}$, which does not have a nontrivial intersection with $\mathcal{H}_K$. By Lemma 4, G is a nontrivial refinement kernel for K. ■

According to Lemma 9, one may expect that every kernel has a nontrivial refinement kernel since in general it should be impossible to impose an inner product on $\mathbb{C}^X$ so that it becomes a RKHS. Our next two results confirm this expectation.

Proposition 10 *If the input space X has a finite cardinality, then a kernel K on X has a nontrivial refinement kernel if and only if $K[X]$ is singular.*

Proof By Lemma 9, it suffices to show that $\mathcal{H}_K = \mathbb{C}^X$ if and only if the matrix $K[X]$ is invertible. Suppose that the cardinality of X is n and $X = \{x_j : j \in \mathbb{N}_n\}$. Assume that $K[X]$ is invertible. Then for each function $\varphi \in \mathbb{C}^X$ there exists a unique vector $[c_j : j \in \mathbb{N}_n] \in \mathbb{C}^n$ such that

$$\sum_{k\in\mathbb{N}_n} c_k K(x_j, x_k) = \varphi(x_j), \quad j \in \mathbb{N}_n,$$

which implies that $\varphi = \sum_{k\in\mathbb{N}_n} c_k K(\cdot, x_k)$. By (4), we have $\varphi \in \mathcal{H}_K$, thereby proving that $\mathcal{H}_K = \mathbb{C}^X$. Conversely, suppose that $\mathcal{H}_K = \mathbb{C}^X$. For each $j \in \mathbb{N}_n$, we introduce a function $\varphi_j \in \mathbb{C}^X$ by setting for each $l \in \mathbb{N}_n$, $\varphi_j(x_l) := \delta_{j,l}$, where δ denotes the Kronecker delta function. By (4) and the assumption that $\mathcal{H}_K = \mathbb{C}^X$, there exists a vector $[c_{j,k} : k \in \mathbb{N}_n] \in \mathbb{C}^n$ such that

$$\sum_{k\in\mathbb{N}_n} c_{j,k} K(x_l, x_k) = \varphi_j(x_l) = \delta_{j,l}, \quad j, l \in \mathbb{N}_n.$$

That is, $[c_{j,k} : j, k \in \mathbb{N}_n]$ is the inverse of the transpose of $K[X]$. Therefore, the matrix $K[X]$ is invertible. The proof is complete. ■

Theorem 11 *If the input space X has an infinite cardinality, then every kernel on it has a nontrivial refinement kernel.*

Proof According to Lemma 9, it suffices to show that there does not exist a kernel K on X such that $\mathcal{H}_K$ contains every function on X. We prove this by contradiction. Assume that there were a kernel K on X such that $\mathcal{H}_K = \mathbb{C}^X$. Since X has a countable subset of distinct points x_n, $n \in \mathbb{N}$, we may define a fixed function $f \in \mathbb{C}^X$ by setting $f(x_j) := j$ for each $j \in \mathbb{N}$ and $f(x) := 0$ for $x \in X \setminus \{x_j : j \in \mathbb{N}\}$. By (3), we would have for each $n \in \mathbb{N}$ that

$$n = |f(x_n)| = |(f, K(\cdot, x_n))_{\mathcal{H}_K}| \leq \|f\|_{\mathcal{H}_K} \|K(\cdot, x_n)\|_{\mathcal{H}_K}. \tag{35}$$

Note that

$$\|K(\cdot, x_n)\|_{\mathcal{H}_K} = \sqrt{K(x_n, x_n)}, \quad n \in \mathbb{N}. \tag{36}$$

Combining (35) and (36) yields that

$$\lim_{n\to\infty} K(x_n, x_n) = +\infty. \tag{37}$$

However, again by (3) for the function $g \in \mathbb{C}^X$ defined by $g(x) := K(x,x)$, $x \in X$, we observe for each $n \in \mathbb{N}$ that

$$\begin{aligned} K(x_n,x_n) &= |g(x_n)| = |(g,K(\cdot,x_n))_{\mathcal{H}_K}| \\ &\leq \|g\|_{\mathcal{H}_K}\|K(\cdot,x_n)\|_{\mathcal{H}_K} = \|g\|_{\mathcal{H}_K}\sqrt{K(x_n,x_n)}. \end{aligned}$$

This implies that

$$K(x_n,x_n) \leq \|g\|^2_{\mathcal{H}_K}, \ \ n \in \mathbb{N},$$

which contradicts (37). The contradiction proves the desired result. ■

In the rest of this section, we show that the refinement process preserves the strictly positive definiteness, the continuity and the universality of the original kernel. A kernel K on X is said to be *strictly positive definite* if for any finite inputs $\mathbf{x} := \{x_j : j \in \mathbb{N}_n\} \subseteq X$ the matrix $K[\mathbf{x}]$ defined by (1) is strictly positive definite. Strictly positive definite kernels are important to the minimum norm interpolation in RKHS.

Proposition 12 *If K is a strictly positive definite kernel on X and G is a refinement kernel for K, then G is also strictly positive definite.*

Proof By Lemma 4, if G is a refinement kernel for K then $G-K$ remains a kernel on X. As a consequence, we have for all $\mathbf{x} := \{x_j : j \in \mathbb{N}_n\} \subseteq X$ that

$$G[\mathbf{x}] = K[\mathbf{x}] + (G-L)[\mathbf{x}].$$

Since $K[\mathbf{x}]$ is strictly positive definite and $(G-L)[\mathbf{x}]$ is positive semi-definite, $G[\mathbf{x}]$ is strictly positive definite. ■

The next kernel property that we consider is continuity. Suppose that the input space X is a topological space. We call a kernel on X a *continuous kernel* if it is at the same time a continuous function on $X \times X$. Given a continuous kernel K on X, can we find a nontrivial refinement kernel G for K that is also continuous? Assuming that X is a metric space with an infinite cardinality, the answer to this question is positive. Recall the Tietze extension theorem in topology (see, for example, Munkres, 2000, page 219), which states that a continuous function defined on a closed subspace of X can be extended to a continuous function on X.

Theorem 13 *If X is a metric space with an infinite cardinality, then every continuous kernel on X has a nontrivial continuous refinement kernel.*

Proof If the topology on X is discrete then any function on X is continuous. In this case, the result holds true by Theorem 11.

Now we suppose that X has an accumulation point x_0. In other words, there exists a sequence of distinct points $x_n \in X$, $n \in \mathbb{N}$ that converges to x_0 and none of the points is the same as x_0. By the arguments used in the proof of Lemma 9, it suffices to prove that there is not a continuous kernel K on X for which $\mathcal{H}_K$ contains all the continuous functions on X. Suppose to the contrary that there is such a kernel K. Then we introduce a sequence of nonnegative numbers by setting

$$c_n := \|K(\cdot,x_n) - K(\cdot,x_{n+1})\|_{\mathcal{H}_K}, \ \ n \in \mathbb{N}.$$

For each $n \in \mathbb{N}$, we define a function φ_n on the closed set $\{x_n, x_{n+1}\}$ as $\varphi_n(x_n) := 1$, $\varphi_n(x_{n+1}) := 0$. The function φ_n is continuous on $\{x_n, x_{n+1}\}$. Therefore, by the Tietze extension theorem, it can be extended to a continuous function on X. As a consequence, for each $n \in \mathbb{N}$, c_n is positive since otherwise we would get by (3) for each $f \in C(X) \subseteq \mathcal{H}_K$ that $f(x_n) = f(x_{n+1})$. By (5), we have that

$$c_n = \sqrt{K(x_n,x_n) + K(x_{n+1},x_{n+1}) - K(x_n,x_{n+1}) - K(x_{n+1},x_n)},\ \ n \in \mathbb{N}.$$

Since K is continuous and x_n converges to x_0, c_n converges to zero as n tends to infinity.

Set $\mathcal{Z} := \{x_n : n \in \mathbb{N}\} \cup \{x_0\}$. Then $\mathcal{Z}$ is a closed subspace of X. We define a continuous function f on $\mathcal{Z}$ by

$$f(x) := \begin{cases} \sqrt{c_{2n-1}}, & x = x_{2n-1},\ n \in \mathbb{N}, \\ 0, & \text{otherwise.} \end{cases}$$

By the Tietze extension theorem, f can be extended to a continuous function on X, which we still denote by f. By the assumption that $C(X) \subseteq \mathcal{H}_K$, $f \in \mathcal{H}_K$. We now obtain by the reproducing property (3) for each $n \in \mathbb{N}$ that

$$\begin{aligned} \sqrt{c_{2n-1}} &= |f(x_{2n-1}) - f(x_{2n})| = |(f, K(\cdot,x_{2n-1}) - K(\cdot,x_{2n}))_{\mathcal{H}_K}| \\ &\le \|f\|_{\mathcal{H}_K} \|K(\cdot,x_{2n-1}) - K(\cdot,x_{2n})\|_{\mathcal{H}_K} = \|f\|_{\mathcal{H}_K} c_{2n-1}. \end{aligned}$$

Thus we get that

$$\|f\|_{\mathcal{H}_K} \ge \frac{1}{\sqrt{c_{2n-1}}},\ \ n \in \mathbb{N},$$

which contradicts the fact that c_n converges to zero. This contradiction implies that $\mathcal{H}_K$ can not contain the space $C(X)$. ■

The result in Theorem 13 remains valid if we only assume that X is a *normal* topological space (see, Munkres, 2000, page 195).

The last kernel property with which we are concerned is universality (Micchelli et al., 2003, 2006; Steinwart, 2001). Suppose that X is a locally compact Hausdorff space. We say that a function $K : X \times X \to \mathbb{C}$ is a *universal kernel* if it is a continuous kernel on X and for all compact subsets $\mathcal{Z} \subseteq X$, $\operatorname{span}\{K(\cdot,x) : x \in \mathcal{Z}\}$ is dense in the Banach space $C(\mathcal{Z})$ of the continuous functions on $\mathcal{Z}$. Universal kernels were extensively studied in Micchelli et al. (2006). They are those kernels that can be used to approximate any continuous target function uniformly on a compact input space.

Proposition 14 *If K is a universal kernel on X, then any continuous refinement kernel for K is universal.*

Proof Suppose that K is a universal kernel on X and G is a continuous refinement kernel for K. Assume that K and G have feature maps $\Phi : X \to \mathcal{W}$ and $\Phi' : X \to \mathcal{W}'$, respectively. By Theorem 4 in Micchelli et al. (2006), G is universal if and only if for all compact $\mathcal{Z} \subseteq X$, $\operatorname{span}\{(\Phi'(\cdot), u')_{\mathcal{W}'} : u' \in \mathcal{W}'\}$ is dense in $C(\mathcal{Z})$. Let $\mathcal{Z}$ be a compact subset of X. Since K is universal, Theorem 4 in Micchelli et al. (2006) ensures that

$$\overline{\operatorname{span}}\{(\Phi(\cdot), u)_{\mathcal{W}} : u \in \mathcal{W}\} = C(\mathcal{Z}). \tag{38}$$

By Theorem 6, there exists a bounded linear operator $T : \mathcal{W}' \to \mathcal{W}$ that satisfies (28). We hence get for each $u \in \mathcal{W}$ that

$$(\Phi(\cdot), u)_{\mathcal{W}} = (T\Phi'(\cdot), u)_{\mathcal{W}} = (\Phi'(\cdot), T^*u)_{\mathcal{W}'}.$$

Therefore, there holds

$$\{(\Phi(\cdot), u)_{\mathcal{W}} : u \in \mathcal{W}\} \subseteq \{(\Phi'(\cdot), u')_{\mathcal{W}'} : u' \in \mathcal{W}'\}.$$

The above inclusion together with (38) proves that G is also universal. ■

By Theorems 11 and 13, it is reasonable to conjecture that there exist nontrivial refinement kernels for most kernels used in machine learning. To verify this conjecture and present concrete examples, we shall discuss refinement kernels for translation invariant kernels and Hilbert-Schmidt kernels in the next two sections.

5. Refinement of Translation Invariant Kernels

In this section, we specify our input space as $\mathbb{R}^d$, $d \in \mathbb{N}$ and investigate refinement kernels for translation invariant kernels on $\mathbb{R}^d$. The presentation of this section is organized into six subsections. We discuss in the first subsection the notion of translation invariant kernels. In Section 5.2 we establish various characterizations of refinement for general translation invariant kernels. We then consider several types of specific translation invariant kernels. Specifically, refinement of B-spline kernels, radial kernels and periodic kernels is studied in Sections 5.3, 5.4 and 5.5, respectively. Finally in Section 5.6, we deal with refinement through an expanding matrix.

5.1 Translation Invariant Kernels

A kernel K on $\mathbb{R}^d$ is said to be *translation invariant* if for all $a \in \mathbb{R}^d$,

$$K(x-a, y-a) = K(x,y), \quad x,y \in \mathbb{R}^d. \tag{39}$$

For each $a \in \mathbb{R}^d$ we introduce the translation operator τ_a by setting for all functions f on $\mathbb{R}^d$

$$\tau_a f := f(\cdot - a).$$

It can be seen by (3) and (39) that a translation invariant kernel K on $\mathbb{R}^d$ satisfies for all $a,b,x,y \in \mathbb{R}^d$ that

$$\tau_a K(\cdot, x) = K(\cdot - a, x) = K(\cdot, x+a) \tag{40}$$

and

$$(\tau_a K(\cdot, y), \tau_b K(\cdot, x))_{\mathcal{H}_K} = (K(\cdot, y+a), K(\cdot, x+b))_{\mathcal{H}_K} = K(x+b, y+a). \tag{41}$$

Recall that $\mathcal{B}(\mathbb{R}^d)$ denotes the set of all the finite positive Borel measures on $\mathbb{R}^d$. It was established by Bochner in Bochner (1959) that K is a continuous translation invariant kernel on $\mathbb{R}^d$ if and only if there exists a $\mu \in \mathcal{B}(\mathbb{R}^d)$ such that

$$K(x,y) = \int_{\mathbb{R}^d} e^{i(x-y,\xi)} d\mu(\xi), \quad x,y \in \mathbb{R}^d,$$

where $(\cdot,\cdot)$ denotes the standard inner product on $\mathbb{R}^d$. This result is referred to as the Bochner theorem. We present below a characterization of translation invariant kernels in terms of their RKHS. To this end, we call a linear operator from a Hilbert space $\mathcal{W}$ to itself an *isomorphism* on $\mathcal{W}$ if it is isomorphic from $\mathcal{W}$ to $\mathcal{W}$.

Proposition 15 *A kernel K on $\mathbb{R}^d$ is translation invariant if and only if for each $a \in \mathbb{R}^d$, τ_a is an isomorphism on $\mathcal{H}_K$.*

Proof Suppose that K is a translation invariant kernel on $\mathbb{R}^d$. Set $a \in \mathbb{R}^d$, $f \in \mathcal{H}_K$ and $\mathcal{H} := \operatorname{span}\{K(\cdot,x) : x \in \mathbb{R}^d\}$. By (4), $\mathcal{H}$ is a dense subspace of $\mathcal{H}_K$. Thus there exists a sequence of functions $f_n \in \mathcal{H}$, $n \in \mathbb{N}$ that converges to f in $\mathcal{H}_K$. By (40) and (41), $\tau_a f_n \in \mathcal{H}$ and $\|\tau_a f_n\|_{\mathcal{H}_K} = \|f_n\|_{\mathcal{H}_K}$ for each $n \in \mathbb{N}$. The latter implies that $\tau_a f_n$ form a Cauchy sequence in $\mathcal{H}_K$. Let g be their limit in $\mathcal{H}_K$. We have that $\|g\|_{\mathcal{H}_K} = \|f\|_{\mathcal{H}_K}$. To prove that τ_a is isometric on $\mathcal{H}_K$, it remains to prove that $g = \tau_a f$. To this end, we verify for each $x \in \mathbb{R}^d$ by (3), (40) and (41) that

$$\begin{aligned} g(x) &= (g, K(\cdot,x))_{\mathcal{H}_K} = \lim_{n\to\infty} (\tau_a f_n, K(\cdot,x))_{\mathcal{H}_K} \\ &= \lim_{n\to\infty} (f_n, K(\cdot,x-a))_{\mathcal{H}_K} = (f, K(\cdot,x-a))_{\mathcal{H}_K} = f(x-a). \end{aligned}$$

Similarly, it can be proved that $\tau_{-a} f_n$ converges to $\tau_{-a} f$, implying that $\tau_{-a} f \in \mathcal{H}_K$. Since $\tau_a \tau_{-a} f = f$, τ_a is surjective from $\mathcal{H}_K$ to $\mathcal{H}_K$. This together with τ_a being isometric shows that it is an isomorphism on $\mathcal{H}_K$.

Conversely, suppose that for each $a \in \mathbb{R}^d$, τ_a is an isomorphism on $\mathcal{H}_K$. This implies that the adjoint operator τ_a^* of τ_a is identified with τ_{-a} (see, Conway, 1990, page 32). It follows for each $x, y \in \mathbb{R}^d$ that $\tau_a K(\cdot,y), \tau_{-a} K(\cdot,x) \in \mathcal{H}_K$ and

$$\begin{aligned} (\tau_a K(\cdot,y), K(\cdot,x))_{\mathcal{H}_K} &= (K(\cdot,y), \tau_{-a} K(\cdot,x))_{\mathcal{H}_K} \\ &= (K(\cdot,y), K(\cdot + a, x))_{\mathcal{H}_K} = K(x, y+a). \end{aligned}$$

On the other hand, we have by (5) that

$$(\tau_a K(\cdot,y), K(\cdot,x))_{\mathcal{H}_K} = (K(\cdot - a, y), K(\cdot,x))_{\mathcal{H}_K} = K(x-a, y).$$

Combining the above two equations, we obtain that $K(x-a,y) = K(x,y+a)$ for all $a,x,y \in \mathbb{R}^d$. Replacing y with $y-a$ yields (39). ■

5.2 Characterizations

Suppose that K is a continuous translation invariant kernel on $\mathbb{R}^d$. We are interested in constructing refinement kernels for K that are continuous and translation invariant as well. Specifically, with a different measure $\mu' \in \mathcal{B}(\mathbb{R}^d)$ we introduce a new kernel

$$G(x,y) := \int_{\mathbb{R}^d} e^{i(x-y,\xi)} d\mu'(\xi), \;\; x,y \in \mathbb{R}^d$$

and characterize G being a refinement kernel for K in terms of a relation between μ and μ'.

Theorem 16 *There holds $\mathcal{H}_K \preceq \mathcal{H}_G$ if and only if $\mu \preceq \mu'$. Moreover, if $\mu \preceq \mu'$ then G is a nontrivial refinement kernel for K if and only if*

$$\mu'(\mathbb{R}^d) - \mu(\mathbb{R}^d) > 0.$$

Proof We prove this result by employing Theorem 8 with $Y := \mathbb{R}^d$. To this end, we introduce a mapping $\phi : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{C}$ by setting $\phi(x,\xi) := e^{i(x,\xi)}$, for $x,\xi \in \mathbb{R}^d$. One can see that K can be represented by

$$K(x,y) = \int_{\mathbb{R}^d} \phi(x,\xi)\overline{\phi(y,\xi)}d\mu(\xi),\ \ x,y \in \mathbb{R}^d$$

and likewise

$$G(x,y) = \int_{\mathbb{R}^d} \phi(x,\xi)\overline{\phi(y,\xi)}d\mu'(\xi),\ \ x,y \in \mathbb{R}^d.$$

Note also that for any $\omega \in \mathcal{B}(\mathbb{R}^d)$, $\operatorname{span}\{\phi(x,\cdot) : x \in \mathbb{R}^d\}$ is dense in $L^2(\mathbb{R}^d,\omega)$. Thus, the result of this theorem is an immediate consequence of Theorem 8. ■

We next characterize the inclusion $\mathcal{H}_K \preceq \mathcal{H}_G$ by using the structure of $\mathcal{H}_K$ and $\mathcal{H}_G$. Let us prepare for this analysis by recalling some basic facts about Borel measures on $\mathbb{R}^d$. Suppose $\nu,\omega \in \mathcal{B}(\mathbb{R}^d)$. If there is a Borel subset $V \subseteq \mathbb{R}^d$ such that for each Borel $U \subseteq \mathbb{R}^d$, $\nu(U) = \nu(U \cap V)$, we say that ν is *concentrated* on V. We call ν a *singular* measure with respect to ω if there exist disjoint Borel subsets U,V of $\mathbb{R}^d$ such that ω is concentrated on U and ν is concentrated on V. The Lebesgue decomposition theorem (see, for example, Rudin, 1987, page 121) asserts that for two measures $\nu,\omega \in \mathcal{B}(\mathbb{R}^d)$, there exist two unique measures $\nu_c,\nu_s \in \mathcal{B}(\mathbb{R}^d)$ with ν_c being absolutely continuous with respect to ω and ν_s being singular with respect to ω such that ν has the Lebesgue decomposition with respect to ω

$$\nu = \nu_c + \nu_s.$$

The Lebesgue decomposition of measures with respect to the Lebesgue measure leads to a decomposition of the corresponding continuous translation invariant kernel. Specifically, for a continuous translation invariant kernel K on $\mathbb{R}^d$, we have the *Lebesgue decomposition*

$$K = K_c + K_s, \tag{42}$$

where

$$K_c(x,y) = \int_{\mathbb{R}^d} e^{i(x-y,\xi)}d\mu_c(\xi),\ \ K_s(x,y) = \int_{\mathbb{R}^d} e^{i(x-y,\xi)}d\mu_s(\xi),\ \ x,y \in \mathbb{R}^d.$$

Likewise, for a continuous translation invariant kernel G on $\mathbb{R}^d$, we also have its Lebesgue decomposition

$$G = G_c + G_s, \tag{43}$$

where

$$G_c(x,y) = \int_{\mathbb{R}^d} e^{i(x-y,\xi)}d\mu'_c(\xi),\ \ G_s(x,y) = \int_{\mathbb{R}^d} e^{i(x-y,\xi)}d\mu'_s(\xi),\ \ x,y \in \mathbb{R}^d.$$

In the above equations, the measures μ_c,μ'_c are absolutely continuous with respect to the Lebesgue measure and μ_s,μ'_s are singular with respect to the Lebesgue measure. By the Radon-Nikodym theorem, there exist nonnegative functions $k,g \in L^1(\mathbb{R}^d)$ such that

$$K_c(x,y) = \int_{\mathbb{R}^d} e^{i(x-y,\xi)}k(\xi)d\xi,\ G_c(x,y) = \int_{\mathbb{R}^d} e^{i(x-y,\xi)}g(\xi)d\xi,\ \ x,y \in \mathbb{R}^d. \tag{44}$$

The next task is to characterize $\mathcal{H}_K \preceq \mathcal{H}_G$ in terms of k, g, μ_s, μ_s'. We start with a simple observation. We associate with each $\psi \in L^1(\mathbb{R}^d)$ a finite Borel measure μ_ψ on $\mathbb{R}^d$ defined on each Borel subset $V \subseteq \mathbb{R}^d$ by

$$\mu_\psi(V) := \int_V \psi(\xi) d\xi.$$

Lemma 17 *If L is a continuous translation invariant kernel on $\mathbb{R}^d$ with a Lebesgue decomposition $L = L_c + L_s$, then $\mathcal{H}_L$ is equal to the orthogonal direct sum of $\mathcal{H}_{L_c}$ and $\mathcal{H}_{L_s}$, namely, $\mathcal{H}_L = \mathcal{H}_{L_c} \bigoplus \mathcal{H}_{L_s}$.*

Proof By Lemma 4, it suffices to show that $\mathcal{H}_{L_c} \cap \mathcal{H}_{L_s} = \{0\}$. We assume that

$$L_c(x,y) = \int_{\mathbb{R}^d} e^{i(x-y,\xi)} l(\xi) d\xi,\ L_s(x,y) = \int_{\mathbb{R}^d} e^{i(x-y,\xi)} d\rho(\xi),\ \ x,y \in \mathbb{R}^d,$$

where $l \in L^1(\mathbb{R}^d)$ is nonnegative and $\rho \in \mathcal{B}(\mathbb{R}^d)$ is singular with respect to the Lebesgue measure. Suppose that $f \in \mathcal{H}_{L_c} \cap \mathcal{H}_{L_s}$. By Lemma 5, there exists $g \in L^2(\mathbb{R}^d, \mu_l)$ and $h \in L^2(\mathbb{R}^d, \rho)$ such that

$$f(x) = \int_{\mathbb{R}^d} e^{i(x,\xi)} g(\xi) l(\xi) d\xi = \int_{\mathbb{R}^d} e^{i(x,\xi)} h(\xi) d\rho(\xi),\ \ x \in \mathbb{R}^d.$$

Define the Borel measure $\mu_{h,\rho}$ on each Borel set $V \subseteq \mathbb{R}^d$ by

$$\mu_{h,\rho}(V) := \int_V h(\xi) d\rho(\xi).$$

By the uniqueness of Fourier transforms (Grafakos, 2004), the two Borel measures μ_{gl}, $\mu_{h,\rho}$ are identical. However, μ_{gl} is absolutely continuous with respect to the Lebesgue measure while $\mu_{h,\rho}$ is singular with respect to the Lebesgue measure. Therefore, we must have $\mu_{gl} = \mu_{h,\rho} = 0$. Consequently, $f = 0$. The proof is complete. ■

The next result allows us to identify the inclusion $\mathcal{H}_K \preceq \mathcal{H}_G$ with two independent inclusions $\mathcal{H}_{K_c} \preceq \mathcal{H}_{G_c}$ and $\mathcal{H}_{K_s} \preceq \mathcal{H}_{G_s}$.

Proposition 18 *Suppose that K and G are continuous translation invariant kernels on $\mathbb{R}^d$ defined by (42) and (43). Then $\mathcal{H}_K \preceq \mathcal{H}_G$ if and only if $\mathcal{H}_{K_c} \preceq \mathcal{H}_{G_c}$ and $\mathcal{H}_{K_s} \preceq \mathcal{H}_{G_s}$.*

Proof Suppose that $\mathcal{H}_{K_c} \preceq \mathcal{H}_{G_c}$ and $\mathcal{H}_{K_s} \preceq \mathcal{H}_{G_s}$. Let f be an arbitrary function in $\mathcal{H}_K$. By Lemma 17, there exists $f_c \in \mathcal{H}_{K_c}$ and $f_s \in \mathcal{H}_{K_s}$ such that $f = f_c + f_s$ and

$$\|f\|^2_{\mathcal{H}_K} = \|f_c\|^2_{\mathcal{H}_{K_c}} + \|f_s\|^2_{\mathcal{H}_{K_s}}.$$

By the assumption, $f_c \in \mathcal{H}_{G_c}$, $f_s \in \mathcal{H}_{G_s}$ and

$$\|f_c\|_{\mathcal{H}_{G_c}} = \|f_c\|_{\mathcal{H}_{K_c}},\ \|f_s\|_{\mathcal{H}_{G_s}} = \|f_s\|_{\mathcal{H}_{K_s}}.$$

Lemma 17 asserts that $\mathcal{H}_G = \mathcal{H}_{G_c} \bigoplus \mathcal{H}_{G_s}$. As a result, we get that $f \in \mathcal{H}_G$ and

$$\|f\|^2_{\mathcal{H}_G} = \|f_c\|^2_{\mathcal{H}_{G_c}} + \|f_s\|^2_{\mathcal{H}_{G_s}} = \|f_c\|^2_{\mathcal{H}_{K_c}} + \|f_s\|^2_{\mathcal{H}_{K_s}} = \|f\|^2_{\mathcal{H}_K}.$$

Therefore, we have proved that $\mathcal{H}_K \preceq \mathcal{H}_G$.

Conversely, we assume that G is a refinement kernel for K. Suppose that $f \in \mathcal{H}_{K_c}$. By Lemma 17, $f \in \mathcal{H}_K$ and $\|f\|_{\mathcal{H}_K} = \|f\|_{\mathcal{H}_{K_c}}$. By the assumption and Lemma 17, there exist two functions $g_c \in \mathcal{H}_{G_c}$ and $g_s \in \mathcal{H}_{G_s}$ such that $f = g_c + g_s$ and

$$\|f\|^2_{\mathcal{H}_K} = \|f\|^2_{\mathcal{H}_G} = \|g_c\|^2_{\mathcal{H}_{G_c}} + \|g_s\|^2_{\mathcal{H}_{G_s}}.$$

To obtain that $f \in \mathcal{H}_{G_c}$ and $\|f\|_{\mathcal{H}_{K_c}} = \|f\|_{\mathcal{H}_{G_c}}$, it suffices to show that $g_s = 0$. Arguments similar to those used in the proof of Lemma 17 serve this purpose. Since f is an arbitrary function in $\mathcal{H}_{K_c}$, we obtain that $\mathcal{H}_{K_c} \preceq \mathcal{H}_{G_c}$. Likewise, one can show that $\mathcal{H}_{K_s} \preceq \mathcal{H}_{G_s}$. The proof is thus complete. ■

We next consider $\mathcal{H}_{K_c} \preceq \mathcal{H}_{G_c}$. For a nonnegative function f on $\mathbb{R}^d$, we let $\Omega_f := \{x \in \mathbb{R}^d : f(x) > 0\}$. We write $k \preceq g$ to mean that $g = k$ almost everywhere on Ω_k with respect to the Lebesgue measure.

Theorem 19 *There holds $\mathcal{H}_{K_c} \preceq \mathcal{H}_{G_c}$ if and only if $k \preceq g$. Moreover, if $k \preceq g$ then G_c is a nontrivial refinement kernel for K_c if and only if*

$$\int_{\mathbb{R}^d} (g(\xi) - k(\xi))d\xi > 0. \tag{45}$$

Proof The proof for this result when G_c has the form $\lambda K_c(2\cdot, 2\cdot)$ for some positive constant λ was provided in Xu and Zhang (2007), Theorem 23. Arguments similar to those in Xu and Zhang (2007) can prove the general result described here. We give a different proof below based on Theorem 16.

By Theorem 16, $\mathcal{H}_{K_c} \preceq \mathcal{H}_{G_c}$ if and only if $\mu_k \ll \mu_g$ and $d\mu_k/d\mu_g$ equals 1 almost everywhere on Ω_k and equals 0 almost everywhere elsewhere. Therefore, $\mathcal{H}_{K_c} \preceq \mathcal{H}_{G_c}$ if and only if for each Borel $V \subseteq \Omega_k$

$$\int_V k(\xi)d\xi = \int_V g(\xi)d\xi.$$

Clearly, the above equation holds for all Borel $V \subseteq \Omega_k$ if and only if $k \preceq g$. The second statement of the result follows from the observation that the right hand side of (45) is equal to $\mu_g(\mathbb{R}^d) - \mu_k(\mathbb{R}^d)$. ■

We are ready to present a characterization of $\mathcal{H}_K \preceq \mathcal{H}_G$ in terms of conditions on k, g, μ_s, μ'_s.

Theorem 20 *Let K and G be continuous translation invariant kernels on $\mathbb{R}^d$ defined by (42) and (43). Then $\mathcal{H}_K \preceq \mathcal{H}_G$ if and only if $k \preceq g$ and $\mu_s \preceq \mu'_s$. The refinement kernel G is nontrivial for K if and only if there holds (45) or $\mu'_s(\mathbb{R}^d) - \mu_s(\mathbb{R}^d) > 0$.*

Proof The result of this theorem follows directly from Proposition 18 and Theorems 19, 16. ■

The next result is a direct consequence of Theorem 19. Suppose that $k \in L^1(\mathbb{R}^d)$ is positive almost everywhere on $\mathbb{R}^d$ and define

$$K(x,y) := \int_{\mathbb{R}^d} e^{i(x-y,\xi)} k(\xi)d\xi, \quad x,y \in \mathbb{R}^d. \tag{46}$$

Corollary 21 *For $k \in L^1(\mathbb{R}^d)$ positive almost everywhere on $\mathbb{R}^d$, define K as in (46). Suppose that G is a continuous translation invariant kernel on $\mathbb{R}^d$ with a Lebesgue decomposition (43). Then $\mathcal{H}_K \preceq \mathcal{H}_G$ if and only if $G_c = K$. The kernel G is a nontrivial refinement kernel for K if and only if $G_c = K$ and $G_s \neq 0$.*

5.3 B-spline Kernels

Our next example is concerned with the B-spline kernel. For a nontrivial compactly supported function $f_0 \in L^2(\mathbb{R}^d)$ such that

$$\overline{f_0(-x)} = f_0(x), \ \ x \in \mathbb{R}^d, \tag{47}$$

we define recursively

$$f_n(x) := \int_{\mathbb{R}^d} f_{n-1}(x-y) f_0(y) dy, \ \ x \in \mathbb{R}^d, \ \ n \in \mathbb{N}.$$

For each odd integer $p \in \mathbb{N}$, we let

$$K(x,y) := f_p(x-y), \ \ x,y \in \mathbb{R}^d. \tag{48}$$

In the next proposition, we show that K is a kernel on $\mathbb{R}^d$ and characterize refinement kernels for K. To this end, we need the Fourier transform $\hat{f}$ of a function $f \in L^1(\mathbb{R}^d)$ defined as

$$\hat{f}(\xi) := \int_{\mathbb{R}^d} f(x) e^{-i(x,\xi)} dx, \ \ \xi \in \mathbb{R}^d.$$

By a standard approximation process (Grafakos, 2004), the Fourier transform can be extended to be a bounded operator on $L^2(\mathbb{R}^d)$.

Proposition 22 *For each odd integer $p \in \mathbb{N}$, K defined by (48) is a kernel on $\mathbb{R}^d$. Moreover, suppose that G is a continuous translation invariant kernel on $\mathbb{R}^d$ with a Lebesgue decomposition (43). Then G is a refinement kernel for K if and only if $G_c = K$.*

Proof Since f_0 is compactly supported, $f_0 \in L^1(\mathbb{R}^d)$. By the Schwartz inequality and by induction, we have that $f_p \in L^1(\mathbb{R}^d)$. By the Fourier transform of convolutions, we know that

$$(f_p)\hat{} = ((f_0)\hat{})^{p+1}.$$

Condition (47) ensures that $(f_0)\hat{}$ is real on $\mathbb{R}^d$. Since p is odd, $(f_p)\hat{}$ is nonnegative. By the Bochner theorem, to prove that K is a kernel on $\mathbb{R}^d$, it suffices to show that $((f_0)\hat{})^{p+1} \in L^1(\mathbb{R}^d)$. This is clear since $((f_0)\hat{})^2 \in L^1(\mathbb{R}^d)$ and $(f_0)\hat{}$ is bounded.

Since f_0 is compactly supported, by the Paley-Wiener theorem (see, for example, Gasquet and Witomski, 1999, page 293), $(f_0)\hat{}$ is real-analytic on $\mathbb{R}^d$. Also, it is nontrivial since f_0 is assumed to be nontrivial. By Corollary 21, to conclude our second statement it suffices to point out the well-known fact that the zeros of a nontrivial real-analytic function on $\mathbb{R}^d$ form a set of Lebesgue measure zero in $\mathbb{R}^d$. ■

A particular example of (48) are the *B-spline kernels* (see, for example, Schölkopf and Smola, 2002, page 98, and the references therein), which are defined as (48) by f_0 that is the characteristic function of a ball in $\mathbb{R}^d$ centered at the origin.

5.4 Radial Kernels

We next turn to the *radial* kernels on $\mathbb{R}^d$. They are kernels of the form

$$K(x,y) := r(\|x-y\|),\ \ x,y \in \mathbb{R}^d, \tag{49}$$

where r is a function on $\mathbb{R}_+ := [0,+\infty)$ and $\|\cdot\|$ denotes the standard Euclidean norm on $\mathbb{R}^d$. It was proved in Schoenberg (1938) that the function K in the form (49) with a continuous function r on $\mathbb{R}_+$ defines a kernel on $\mathbb{R}^d$ for all $d \in \mathbb{N}$ if and only if there exists some $\mu \in \mathcal{B}(\mathbb{R}_+)$ such that

$$r(t) := \int_{\mathbb{R}_+} e^{-\sigma t^2} d\mu(\sigma),\ \ t \in \mathbb{R}_+. \tag{50}$$

Theorem 23 *Suppose that K is a nontrivial radial kernel defined by (49), (50) with $\mu(\{0\}) = 0$. Then a continuous translation invariant kernel G on $\mathbb{R}^d$ with a Lebesgue decomposition (43) is a refinement kernel for K if and only if $G_c = K$.*

Proof We prove this theorem by applying Corollary 21 with identifying the function $k \in L^1(\mathbb{R}^d)$ that is positive almost everywhere.

Recalling that for all $\sigma > 0$ there holds

$$\exp\left(-\sigma\|x-y\|^2\right) = \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}\left(\frac{\pi}{\sigma}\right)^{d/2} e^{i(x-y,\xi)} e^{-\frac{\|\xi\|^2}{4\sigma}} d\xi,\ \ x,y \in \mathbb{R}^d,$$

by the hypothesis that $\mu(\{0\}) = 0$, we have for all $x,y \in \mathbb{R}^d$ that

$$K(x,y) = \int_{(0,+\infty)} \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}\left(\frac{\pi}{\sigma}\right)^{d/2} e^{i(x-y,\xi)} e^{-\frac{\|\xi\|^2}{4\sigma}} d\xi d\mu(\sigma).$$

Define

$$k(\xi) := \int_{(0,+\infty)} \frac{1}{(2\pi)^d}\left(\frac{\pi}{\sigma}\right)^{d/2} e^{-\frac{\|\xi\|^2}{4\sigma}} d\mu(\sigma),\ \ \xi \in \mathbb{R}^d.$$

It can be verified by the Fubini theorem (see, Rudin, 1987, page 164) that

$$\int_{\mathbb{R}^d} k(\xi) d\xi = \int_{(0,+\infty)} d\mu(\sigma)\int_{\mathbb{R}^d}\frac{1}{(2\pi)^d}\left(\frac{\pi}{\sigma}\right)^{d/2} e^{-\frac{\|\xi\|^2}{4\sigma}} d\xi = \int_{(0,+\infty)} d\mu(\sigma) = \mu(\mathbb{R}_+).$$

Therefore, k is a nontrivial function in $L^1(\mathbb{R}^d)$. Again, by the Fubini theorem we get that

$$K(x,y) = \int_{\mathbb{R}^d} e^{i(x-y,\xi)} k(\xi) d\xi,\ \ x,y \in \mathbb{R}^d.$$

Thus, to conclude the result of this theorem by Corollary 21, it remains to show that k is positive almost everywhere on $\mathbb{R}^d$. To this end, we observe that the function

$$\varphi(t) := \int_{(0,+\infty)} \frac{1}{(2\pi)^d}\left(\frac{\pi}{\sigma}\right)^{d/2} e^{-\frac{t}{4\sigma}} d\mu(\sigma),\ \ t > 0$$

belongs to $C^\infty(0,+\infty)$ and satisfies for each nonnegative integer j that $(-1)^j\varphi^{(j)}(t) > 0$, for $t > 0$. In other words, φ is *completely monotonic* in $(0,+\infty)$ and is hence real-analytic on the interval (see, Widder, 1941, pages 145–146). Consequently, $k = \varphi(\|\cdot\|^2)$ is real-analytic on $\mathbb{R}^d \setminus \{0\}$. This together with $\|k\|_{L^1(\mathbb{R}^d)} > 0$ proves that k is almost everywhere positive on $\mathbb{R}^d$. The proof is complete. ■

We remark that if $\mu(\{0\}) > 0$ then K is the sum of a constant kernel and a radial kernel satisfying the hypothesis of the last theorem. Note that a constant kernel is defined by a singular Borel measure. Therefore, by Theorems 20 and 23, a continuous kernel G with a Lebesgue decomposition (43) is a refinement kernel for K if and only if $G_c = K - \mu(\{0\})$ and $\mu_s'(\{0\}) = \mu(\{0\})$ since μ_s' and μ must agree at the origin.

As a direct consequence of Theorem 23, we have the following result about the Gaussian kernels

$$\mathcal{G}_\sigma(x,y) := \exp\left(-\sigma\|x-y\|^2\right), \ \ x,y \in \mathbb{R}^d, \ \ \sigma > 0.$$

Corollary 24 *A continuous translation invariant kernel G on $\mathbb{R}^d$ with a Lebesgue decomposition (43) is a refinement kernel for the Gaussian kernel $\mathcal{G}_\sigma$ if and only if $G_c = \mathcal{G}_\sigma$. It is a nontrivial refinement kernel for $\mathcal{G}_\sigma$ if and only if $G_c = \mathcal{G}_\sigma$ and $G_s \neq 0$.*

The corollary above suggests that we may refine the Gaussian kernel $\mathcal{G}_\sigma$ by adding to it a kernel G_s defined by a singular measure on $\mathbb{R}^d$. The RKHS for a Gaussian kernel has been well understood (see, for example, Walder et al., 2006). In particular, we can see by Lemma 5 that

$$\mathcal{H}_{\mathcal{G}_\sigma} := \left\{ f \in L^2(\mathbb{R}^d) : \int_{\mathbb{R}^d} |\hat{f}(\xi)|^2 e^{\frac{\|\xi\|^2}{4\sigma}} d\xi < +\infty \right\}, \tag{51}$$

and the inner product on $\mathcal{H}_{\mathcal{G}_\sigma}$ is given as

$$(f,g)_{\mathcal{H}_{\mathcal{G}_\sigma}} = \frac{1}{(2\pi)^d}\left(\frac{\sigma}{\pi}\right)^{d/2} \int_{\mathbb{R}^d} \hat{f}(\xi)\overline{\hat{g}(\xi)} e^{\frac{\|\xi\|^2}{4\sigma}} d\xi.$$

By (51), $\mathcal{H}_{\mathcal{G}_\sigma} \subseteq \mathcal{H}_{\mathcal{G}_{\sigma'}}$ for $\sigma < \sigma'$. However, since functions with a continuous compactly supported Fourier transform are contained in $\mathcal{H}_{\mathcal{G}_\sigma}$ and are dense in $\mathcal{H}_{\mathcal{G}_{\sigma'}}$, $\mathcal{H}_{\mathcal{G}_\sigma}$ is dense in $\mathcal{H}_{\mathcal{G}_{\sigma'}}$. But, $\mathcal{H}_{\mathcal{G}_\sigma}$ is not closed under the norm of $\mathcal{H}_{\mathcal{G}_{\sigma'}}$. Therefore, there does not exist σ' with $\sigma < \sigma'$ such that $\mathcal{G}_{\sigma'}$ is a refinement kernel for $\mathcal{G}_\sigma$.

5.5 Periodic Kernels

We now investigate kernels defined by continuous periodic functions and their refinement. For this purpose, we recall the Fourier coefficients of a function $f \in L^2([0,2\pi]^d)$ which are defined by setting for each $n \in \mathbb{Z}^d$

$$c_n(f) := \frac{1}{(2\pi)^d} \int_{[0,2\pi]^d} f(x) e^{-i(n,x)} dx.$$

A function f on $\mathbb{R}^d$ is called 2π-*periodic* if for all $n \in \mathbb{Z}^d$, $f = f(\cdot + 2\pi n)$.

Proposition 25 *Let f be a continuous 2π-periodic function on $\mathbb{R}^d$. Then $K(x,y) := f(x-y)$ defines a kernel on $\mathbb{R}^d$ if and only if $\mathbf{f} := [c_n(f) : n \in \mathbb{Z}^d] \in \ell^1(\mathbb{Z}^d)$ and $\mathbf{f} \geq 0$.*

Proof If $\mathbf{f} \in \ell^1(\mathbb{Z}^d)$ and $\mathbf{f} \geq 0$, then we define the Borel measure μ that is supported on $\mathbb{Z}^d$ with $\mu(n) := c_n(f)$, $n \in \mathbb{Z}^d$. By the Fourier series expansion of f, we get that

$$K(x,y) = f(x-y) = \sum_{n\in\mathbb{Z}^d} c_n(f) e^{i(n,x-y)} = \int_{\mathbb{R}^d} e^{i(x-y,\xi)} d\mu(\xi), \ \ x,y \in \mathbb{R}^d.$$

By the Bochner theorem, K is a kernel on $\mathbb{R}^d$.

Conversely, if K is a kernel on $\mathbb{R}^d$, then again by the Bochner theorem, there is a $\mu \in \mathcal{B}(\mathbb{R}^d)$ such that

$$f(x) = \int_{\mathbb{R}^d} e^{i(x,\xi)} d\mu(\xi), \ \ x \in \mathbb{R}^d. \tag{52}$$

Since f is 2π-periodic, we have for each $n \in \mathbb{Z}^d$ that

$$\int_{\mathbb{R}^d} e^{i(x,\xi)} d\mu(\xi) = \int_{\mathbb{R}^d} e^{i(x,\xi)} e^{i2\pi(n,\xi)} d\mu(\xi), \ \ x \in \mathbb{R}^d.$$

By the uniqueness of Fourier transforms, there holds for almost every $\xi \in \mathbb{R}^d$ with respect to μ that $e^{i2\pi(n,\xi)} = 1$, for $n \in \mathbb{Z}^d$. Note that this equation holds for all $n \in \mathbb{Z}^d$ if and only if $\xi \in \mathbb{Z}^d$. Therefore, $\mu(\mathbb{R}^d \setminus \mathbb{Z}^d) = 0$. Consequently, by (52) we obtain that

$$f(x) = \sum_{n\in\mathbb{Z}^d} \mu(\{n\}) e^{i(n,x)}, \ \ x \in \mathbb{R}^d.$$

It is implied that $c_n(f) = \mu(\{n\})$, $n \in \mathbb{Z}^d$. Since μ is finite and positive, $\mathbf{f} \in \ell^1(\mathbb{Z}^d)$ and $\mathbf{f} \geq 0$. ■

By the last proposition, for $c := [c_n : n \in \mathbb{Z}^d] \in \ell^1(\mathbb{Z}^d)$ with $c_n \geq 0$ for each $n \in \mathbb{Z}^d$, we introduce kernel

$$f_c(x-y) := \sum_{n\in\mathbb{Z}^d} c_n e^{i(n,x-y)}, \ \ x,y \in \mathbb{R}^d, \tag{53}$$

and set $\Omega_c := \{n \in \mathbb{Z}^d : c_n > 0\}$. The function $c \in \ell^1(\mathbb{Z}^d)$ will be viewed at the same time as a Borel measure on $\mathbb{Z}^d$ whose measure on $n \in \mathbb{Z}^d$ is defined to be c_n.

Proposition 26 *Suppose that $a, b \in \ell^1(\mathbb{Z}^d)$ with $a_n, b_n \geq 0$ for each $n \in \mathbb{Z}^d$ and define f_a, f_b as in (53). Then $\mathcal{H}_{f_a} \preceq \mathcal{H}_{f_b}$ if and only if*

$$\Omega_a \subseteq \Omega_b \ \ \textit{and for all} \ \ n \in \Omega_a, \ \ a_n = b_n. \tag{54}$$

If $\mathcal{H}_{f_a} \preceq \mathcal{H}_{f_b}$ then f_b is a nontrivial refinement kernel for f_a if and only if Ω_a is a proper subset of Ω_b.

Proof By Theorem 20, $\mathcal{H}_{f_a} \preceq \mathcal{H}_{f_b}$ if and only if $a \preceq b$. We now show that $a \preceq b$ is equivalent to (54).

Suppose that condition (54) holds. By $\Omega_a \subseteq \Omega_b$ we have that $a \ll b$. The derivative $c := da/db$ is given by

$$c_n := \begin{cases} \dfrac{a_n}{b_n}, & n \in \Omega_a, \\ 0, & \text{otherwise}. \end{cases} \tag{55}$$

Therefore, $a \preceq b$.

Conversely, suppose that $a \preceq b$. Since $a \ll b$, we have $\Omega_a \subseteq \Omega_b$. Also, since the derivative $c := da/db$ given by (55) is equal to 1 or 0 almost everywhere with respect to b, we have $a_n = b_n$ for each $n \in \Omega_a$. Hence, (54) holds. ■

5.6 Refinement via an Expanding Matrix

To close this section, we consider a special refinement process in the sense of refinable kernels introduced in Xu and Zhang (2007). The translation invariant kernels K_c defined by (44) via a nonnegative function $k \in L^1(\mathbb{R}^d)$ are of special interest. We call them translation invariant kernels of continuous type. Next, we consider updating kernels of this type through an expanding matrix. Let D be a $d \times d$ real matrix with determinant $\det D$ bigger than 1. Such a matrix is called expanding. Refinable functions with respect to an expanding matrix and the corresponding wavelets were studied by many authors (see, for example, Chen et al., 2003, 2007; Daubechies, 1992; Goodman and Lee, 1994; Jia, 1999; Jia et al., 1999; Micchelli and Sauer, 1997; Micchelli and Xu, 1994; Wang, 2002, and the references cited therein). In particular, the interesting relation between wavelets and tiling was investigated in Wang (2002). For a continuous translation invariant kernel K_c defined by (44), we consider a refinement kernel G_c having the form

$$G_c(x,y) = \lambda K_c(Dx, Dy), \quad x,y \in \mathbb{R}^d. \tag{56}$$

We are interested in characterizing $\mathcal{H}_{K_c} \preceq \mathcal{H}_{G_c}$ in terms of k, λ and D. A special case of this problem when D is the dilation matrix $2I$ was studied in Xu and Zhang (2007).

Theorem 27 *Suppose that K_c is defined by (44) via a nonnegative $k \in L^1(\mathbb{R}^d)$ and G_c is given by (56). Then, $\mathcal{H}_{K_c} \preceq \mathcal{H}_{G_c}$ if and only if for almost every $\xi \in \Omega_k$,*

$$k(\xi) = \frac{\lambda}{\det D} k((D^T)^{-1}\xi).$$

If $\mathcal{H}_{K_c} \preceq \mathcal{H}_{G_c}$ then G_c is a nontrivial refinement kernel for K_c if and only if

$$\int_{\Omega_k \setminus (D^T)^{-1}\Omega_k} k(\xi) d\xi > 0. \tag{57}$$

Proof Through a change of variables, we obtain from (56) that

$$G_c(x,y) = \frac{\lambda}{\det D} \int_{\mathbb{R}^d} e^{i(x-y,\xi)} k((D^T)^{-1}\xi) d\xi, \quad x,y \in \mathbb{R}^d.$$

We then identify the nonnegative function $g \in L^1(\mathbb{R}^d)$ in the definition (44) of the translation invariant kernel G_c of continuous type as follows

$$g(\xi) = \frac{\lambda}{\det D} k((D^T)^{-1}\xi), \quad \xi \in \mathbb{R}^d.$$

The first statement of this theorem now follows directly from the first statement of Theorem 19.

If $\mathcal{H}_{K_c} \preceq \mathcal{H}_{G_c}$, by the first statement of this theorem, we have that

$$\int_{\mathbb{R}^d} g(\xi)d\xi = \int_{\mathbb{R}^d} k(\xi)d\xi + \int_{\mathbb{R}^d\setminus\Omega_k} \frac{\lambda}{\det D} k((D^T)^{-1}\xi)d\xi.$$

Using this identity, we observe that inequality (45) is equivalent to inequality

$$\int_{\mathbb{R}^d\setminus\Omega_k} k((D^T)^{-1}\xi)d\xi > 0.$$

By a change of variables, we find that the inequality above is equivalent to inequality (57). ■

Along this direction, we present a corollary to Proposition 26.

Corollary 28 *Let D be an invertible matrix in $\mathbb{Z}^{d\times d}$, f_a a kernel defined as in (53) and G defined as $G(x,y) = \lambda f_a(Dx - Dy)$ for some positive constant λ. Then G is a refinement kernel for f_a if and only if $\Omega_a \subseteq D^T\Omega_a$ and for each $n \in \Omega_a$, $a_n = \lambda a_{(D^T)^{-1}n}$.*

We shall see in the next section that Proposition 26 and Corollary 28 are special instances of refinement of Hilbert-Schmidt kernels.

6. Refinement of Hilbert-Schmidt Kernels

We characterize in this section refinement kernels for two types of Hilbert-Schmidt kernels. As special examples, we study refinement of the Bergman kernels, the Szegö kernels, the Schoenberg kernels, and kernels having finite dimensional feature spaces.

Let a be a nonnegative function on $\mathbb{N}$ and set $a_n := a(n)$, $n \in \mathbb{N}$. We denote by $\ell^2_a(\mathbb{N})$ the set of functions f on $\mathbb{N}$ such that $\sum_{n\in\mathbb{N}} a_n |f_n|^2 < +\infty$. It is a Hilbert space with the inner product

$$(f,g)_{\ell^2_a(\mathbb{N})} := \sum_{n\in\mathbb{N}} a_n f_n \overline{g_n}, \quad f,g \in \ell^2_a(\mathbb{N}).$$

Suppose that we have a sequence of functions ϕ_n on the input space X, $n \in \mathbb{N}$ such that for each $x \in X$ the function $\Phi(x)$ on $\mathbb{N}$ defined as

$$\Phi(x)(n) := \phi_n(x), \quad n \in \mathbb{N} \tag{58}$$

belongs to $\ell^2_a(\mathbb{N})$. The *Hilbert-Schmidt kernel K_a* associated with a is given as

$$K_a(x,y) := (\Phi(x), \Phi(y))_{\ell^2_a(\mathbb{N})} = \sum_{n\in\mathbb{N}} a_n \phi_n(x)\overline{\phi_n(y)}, \quad x,y \in X. \tag{59}$$

The Mercer theorem (see, for example, Cucker and Smale, 2002; Hochstadt, 1973; Mercer, 1909; Sun, 2005) in the theory of reproducing kernels indicates that (59) represents a large class of kernels. Hilbert-Schmidt kernels are a key element of recent studies Opfer (2006) and Rakotomamonjy and Canu (2005).

The support of a function a on $\mathbb{N}$, denoted as $\operatorname{supp} a$, is the set of $n \in \mathbb{N}$ for which $a_n \neq 0$. Let a, b be two nonnegative functions on $\mathbb{N}$. Set $c := \max\{a, b\}$ and assume that the sequence ϕ_n satisfies for each $x \in X$ that $\Phi(x) \in \ell^2_c(\mathbb{N})$ and

$$\overline{\operatorname{span}}\{\Phi(x) : x \in X\} = \ell^2_c(\mathbb{N}).$$

We shall consider the inclusion $\mathcal{H}_{K_a} \preceq \mathcal{H}_{K_b}$. For this purpose, we make the convention that whenever we write $a \preceq b$ for two functions a,b on $\mathbb{N}$, it means that $\operatorname{supp} a \subseteq \operatorname{supp} b$ and $a_n = b_n$ for each $n \in \operatorname{supp} a$.

Theorem 29 *There holds $\mathcal{H}_{K_a} \preceq \mathcal{H}_{K_b}$ if and only if $a \preceq b$. Moreover, if $a \preceq b$ then K_b is a nontrivial refinement kernel for K_a if and only if* $\operatorname{supp} a$ *is a proper subset of* $\operatorname{supp} b$.

Proof This theorem is proved by using Theorem 8 with an identification of measures μ and ν. We introduce three nonnegative functions $\tilde{a}, \tilde{b}, \tilde{c}$ in $\ell^1(\mathbb{N})$ by setting for $n \in \mathbb{N}$

$$\tilde{a}_n := \begin{cases} \frac{a_n}{n^2 c_n}, & n \in \operatorname{supp} a, \\ 0, & \text{otherwise}, \end{cases} \quad \tilde{b}_n := \begin{cases} \frac{b_n}{n^2 c_n}, & n \in \operatorname{supp} b, \\ 0, & \text{otherwise}, \end{cases}$$

and

$$\tilde{c}_n := \begin{cases} \frac{1}{n^2}, & n \in \operatorname{supp} c, \\ 0, & \text{otherwise}. \end{cases}$$

We also define a function $\phi : X \times \mathbb{N} \to \mathbb{C}$ by

$$\phi(x,n) := n\sqrt{c_n}\phi_n(x), \ \ x \in X, \ n \in \mathbb{N}.$$

It is observed that $\tilde{c} = (\tilde{a} + \tilde{b})/2 + |\tilde{a} - \tilde{b}|/2$, $\phi(x,\cdot) \in \ell^2_{\tilde{c}}(\mathbb{N})$ for all $x \in X$ and $\operatorname{span}\{\phi(x,\cdot) : x \in X\}$ is dense in $\ell^2_{\tilde{c}}(\mathbb{N})$. Moreover,

$$K_a(x,y) = (\phi(x,\cdot),\phi(y,\cdot))_{\ell^2_{\tilde{a}}(\mathbb{N})}, \ K_b(x,y) = (\phi(x,\cdot),\phi(y,\cdot))_{\ell^2_{\tilde{b}}(\mathbb{N})}, \ \ x,y \in X.$$

Let $Y := \mathbb{N}$. We identify measures $\mu, \nu \in \mathcal{B}(Y)$ with $\tilde{a}$ and $\tilde{b}$, respectively, such that

$$K_a(x,y) = \int_Y \phi(x,\xi)\overline{\phi(y,\xi)}d\mu(\xi), \ \ K_b(x,y) = \int_Y \phi(x,\xi)\overline{\phi(y,\xi)}d\nu(\xi), \ \ x,y \in X.$$

Note that $\mu \preceq \nu$ is equivalent to $a \preceq b$. The result of this theorem now follows immediately from Theorem 8. ■

One can see that Proposition 26 may be viewed as a corollary of Theorem 29. We next present two more concrete applications of Theorem 29.

For the first example, we assume $R \in (0, +\infty]$ and specify our input space X to be $\{z \in \mathbb{C} : |z| < R^{1/2}\}$. Here we make the convention that if $R = +\infty$ then $X := \mathbb{C}$. For two nonnegative functions a, b defined on $\mathbb{N}$ satisfying

$$\max\left\{\limsup_{n\to\infty} \sqrt[n]{a_n}, \ \limsup_{n\to\infty} \sqrt[n]{b_n}\right\} \leq \frac{1}{R}, \tag{60}$$

we define the kernels

$$\mathcal{K}_a(\xi,\eta) := \sum_{n\in\mathbb{N}} a_n \xi^{n-1}\overline{\eta}^{n-1}, \ \mathcal{K}_b(\xi,\eta) := \sum_{n\in\mathbb{N}} b_n \xi^{n-1}\overline{\eta}^{n-1}, \ \ \xi,\eta \in X. \tag{61}$$

Classical kernels such as the Bergman kernels and the Szegö kernels (see, for example, Saitoh, 1988) have the above form.

Proposition 30 *Suppose that a, b are nonnegative functions defined on $\mathbb{N}$ satisfying (60) and kernels $\mathcal{K}_a, \mathcal{K}_b$ are defined in (61). Then $\mathcal{H}_{\mathcal{K}_a} \preceq \mathcal{H}_{\mathcal{K}_b}$ if and only if $a \preceq b$, and the refinement is nontrivial if and only if* $\operatorname{supp} a$ *is a proper subset of* $\operatorname{supp} b$.

Proof We define a sequence of functions ϕ_n, $n \in \mathbb{N}$ on the input space X by setting

$$\phi_n(\xi) := \xi^{n-1}, \ \ \xi \in X.$$

Let $c := \max\{a, b\}$. Condition (60) ensures that Φ defined by (58) with the ϕ_n defined above satisfies the condition that $\Phi(\xi) \in \ell_c^2(\mathbb{N})$ for each $\xi \in X$. It is clear that $\operatorname{span}\{\Phi(\xi) : \xi \in X\}$ is dense in $\ell_c^2(\mathbb{N})$. The result of this proposition follows directly from Theorem 29. ■

Our second example concerns the *Schoenberg kernels* (Schoenberg, 1942) on the unit sphere $\mathbb{S}^d$ in $\mathbb{R}^{d+1}$. We shall need the ultraspherical polynomials P_n^d, $n \in \mathbb{Z}_+$. When $d = 1$, P_n^1 is the Chebyshev polynomial of degree n (Rivlin, 1990) and for $d > 1$, P_n^d is determined by

$$\frac{1}{(1-2zt+z^2)^{(d-1)/2}} = \sum_{n\in\mathbb{Z}_+} P_n^d(t) z^n, \ \ |z| < 1, \ t \in [-1, 1].$$

For a nonnegative function h defined on $\mathbb{N}$ satisfying the conditions

$$\sum_{n\in\mathbb{N}} h_n P_{n-1}^d(1) < +\infty, \tag{62}$$

we introduce a Schoenberg kernel on $\mathbb{S}^d$ by setting

$$\mathcal{S}_h(x, y) := \sum_{n\in\mathbb{N}} h_n P_{n-1}^d((x, y)), \ \ x, y \in \mathbb{S}^d, \tag{63}$$

where $(\cdot, \cdot)$ denotes the inner product on $\mathbb{R}^{d+1}$.

Theorem 31 *Suppose that a and b are two nonnegative functions defined on $\mathbb{N}$ satisfying the condition (62) and $\mathcal{S}_a$ and $\mathcal{S}_b$ are the corresponding Schoenberg kernels on $\mathbb{S}^d$ defined as in (63). Then, $\mathcal{H}_{\mathcal{S}_a} \preceq \mathcal{H}_{\mathcal{S}_b}$ if and only if $a \preceq b$. If $a \preceq b$ then $\mathcal{S}_b$ is a nontrivial refinement kernel for $\mathcal{S}_a$ if and only if* $\operatorname{supp} a$ *is a proper subset of* $\operatorname{supp} b$.

Proof We shall write the kernels $\mathcal{S}_a, \mathcal{S}_b$ in the form of Hilbert-Schmidt kernels and then apply Theorem 29. To this end, we recall some basic facts of *spherical harmonics* (Stein and Weiss, 1971). For each $n \in \mathbb{Z}_+$ we let $\mathcal{H}_n$ be the set of all homogeneous harmonic polynomials of total degree n on $\mathbb{R}^{d+1}$ restricted to $\mathbb{S}^d$. We consider $\mathcal{H}_n$ as a subspace of $L^2(\mathbb{S}^d, \omega)$ where ω is the Lebesgue measure on $\mathbb{S}^d$. Let d_n denote the dimension of $\mathcal{H}_n$ and $\{Y_j^n : j \in \mathbb{N}_{d_n}\}$ an orthonormal basis for $\mathcal{H}_n$. If $n \neq n'$ then $\mathcal{H}_n$ is orthogonal to $\mathcal{H}_{n'}$ (Stein and Weiss, 1971). For each $n \in \mathbb{Z}_+$, there exists a positive constant c_n such that

$$P_n^d((x, y)) = c_n \sum_{j\in\mathbb{N}_{d_n}} Y_j^n(x) Y_j^n(y), \ \ x, y \in \mathbb{S}^d. \tag{64}$$

By Equations (63) and (64), we have that

$$\mathcal{S}_a(x, y) = \sum_{n\in\mathbb{N}} a_n c_{n-1} \sum_{j\in\mathbb{N}_{d_{n-1}}} Y_j^{n-1}(x) Y_j^{n-1}(y), \ \ x, y \in \mathbb{S}^d$$

and

$$\mathcal{S}_b(x,y)=\sum_{n\in\mathbb{N}} b_n c_{n-1}\sum_{j\in\mathbb{N}_{d_{n-1}}} Y_j^{n-1}(x)Y_j^{n-1}(y),\ \ x,y\in\mathbb{S}^d.$$

The result of this theorem hence follows immediately from Theorem 29 and the orthogonality of Y_j^n. ■

We now return to general Hilbert-Schmidt kernels defined by a sequence of functions ϕ_n on X and investigate the case when ϕ_n's are coupled. We shall work in the Hilbert space $\ell^2(\mathbb{N})$ under the assumption that $\Phi(x)\in\ell^2(\mathbb{N})$ for each $x\in X$ and

$$\overline{\text{span}}\{\Phi(x):x\in X\}=\ell^2(\mathbb{N}). \tag{65}$$

For each bounded, positive, and self-adjoint linear operator C on $\ell^2(\mathbb{N})$, we denote by $\ell^2_C(\mathbb{N})$ the Hilbert space completed upon the linear space $\ell^2(\mathbb{N})$ under the inner product

$$(u,v)_{\ell^2_C(\mathbb{N})}:=(Cu,v)_{\ell^2(\mathbb{N})},\ \ u,v\in\ell^2(\mathbb{N}).$$

Note that $\ell^2_C(\mathbb{N})$ is a Hilbert space of equivalent classes. In other words, two elements $u,v\in\ell^2(\mathbb{N})$ are identical in $\ell^2_C(\mathbb{N})$ if and only if $u-v\in\ker C:=\{x\in\ell^2(\mathbb{N}):Cx=0\}$. For two bounded, positive and self-adjoint linear operators A,B from $\ell^2(\mathbb{N})$ to itself, we introduce two kernels by setting

$$K_A(x,y):=(\Phi(x),\Phi(y))_{\ell^2_A(\mathbb{N})},\ K_B(x,y):=(\Phi(x),\Phi(y))_{\ell^2_B(\mathbb{N})},\ \ x,y\in X.$$

Before delving into conditions equivalent to $\mathcal{H}_{K_A}\preceq\mathcal{H}_{K_B}$, we review necessary results from functional analysis. For each bounded, positive and self-adjoint linear operator C on $\ell^2(\mathbb{N})$, we let $P_{C,\perp}$ denote the orthogonal projection from $\ell^2(\mathbb{N})$ to the orthogonal complement $(\ker C)^\perp$ of $\ker C$. Note that $u\in\ell^2(\mathbb{N})$ satisfies $(Cu,u)_{\ell^2(\mathbb{N})}=0$ if and only if $u\in\ker C$. Moreover, for each $u\in\ell^2(\mathbb{N})$ there holds that

$$(Cu,u)_{\ell^2(\mathbb{N})}=(CP_{C,\perp}u,P_{C,\perp}u)_{\ell^2(\mathbb{N})}.$$

Thus C is a bijective mapping from $(\ker C)^\perp$ to $\operatorname{ran}C$. We denote by $\tilde{C}^{-1}$ its inverse from $\operatorname{ran}C$ to $(\ker C)^\perp$.

Our characterization of $\mathcal{H}_{K_A}\preceq\mathcal{H}_{K_B}$ is as follows.

Theorem 32 *Suppose that A,B are bounded, positive and self-adjoint linear operators from $\ell^2(\mathbb{N})$ to itself. Then, $\mathcal{H}_{K_A}\preceq\mathcal{H}_{K_B}$ if and only if $\operatorname{ran}A\subseteq\operatorname{ran}B$ and for each $v\in\operatorname{ran}A$,*

$$P_{A,\perp}\tilde{B}^{-1}v=\tilde{A}^{-1}v. \tag{66}$$

Proof By Lemma 5, $\mathcal{H}_{K_A}\preceq\mathcal{H}_{K_B}$ if and only if for each $u\in\ell^2(\mathbb{N})$ there exists $v\in\ell^2(\mathbb{N})$ such that

$$(\Phi(x),Au)_{\ell^2(\mathbb{N})}=(\Phi(x),Bv)_{\ell^2(\mathbb{N})},\ \ x\in X \tag{67}$$

and

$$(Au,u)_{\ell^2(\mathbb{N})}=(Bv,v)_{\ell^2(\mathbb{N})}. \tag{68}$$

By the density assumption (65), (67) is equivalent to that

$$Au=Bv. \tag{69}$$

Suppose that $\mathcal{H}_{K_A} \preceq \mathcal{H}_{K_B}$. Then for each $u \in \ell^2(\mathbb{N})$ there exists $v \in \ell^2(\mathbb{N})$ satisfying (68) and (69). By (69), we have that $\operatorname{ran} A \subseteq \operatorname{ran} B$. Let $u \in (\ker A)^\perp$. Then v in (69) can be taken as $\tilde{B}^{-1}Au$. Similarly, if we choose $u' \in (\ker A)^\perp$ and let $v' := \tilde{B}^{-1}Au'$ then (67) and (69) hold true with u, v replaced by u', v'. Since $\mathcal{H}_{K_A}$ is a subspace of $\mathcal{H}_{K_B}$, we have

$$\begin{aligned}(Au', u)_{\ell^2(\mathbb{N})} &= ((\Phi(x), Au)_{\ell^2(\mathbb{N})}, (\Phi(x), Au')_{\ell^2(\mathbb{N})})_{\mathcal{H}_{K_A}}\\ &= ((\Phi(x), Bv)_{\ell^2(\mathbb{N})}, (\Phi(x), Bv')_{\ell^2(\mathbb{N})})_{\mathcal{H}_{K_B}}\\ &= (Bv', v)_{\ell^2(\mathbb{N})}.\end{aligned}$$

By the above equation and (69), we get that

$$(Au', P_{A,\perp}\tilde{B}^{-1}Au)_{\ell^2(\mathbb{N})} = (Au', \tilde{B}^{-1}Au)_{\ell^2(\mathbb{N})} = (Bv', v)_{\ell^2(\mathbb{N})} = (Au', u)_{\ell^2(\mathbb{N})}.$$

Note that the above equation is true for all $u, u' \in (\ker A)^\perp$. Thus there must hold

$$P_{A,\perp}\tilde{B}^{-1}Au = u = \tilde{A}^{-1}Au.$$

Since A is surjective from $(\ker A)^\perp$ onto $\operatorname{ran} A$, we obtain (66) for each $v \in \operatorname{ran} A$.

Conversely, suppose that $\operatorname{ran} A \subseteq \operatorname{ran} B$ and there holds for each $v \in \operatorname{ran} A$ Equation (66). Therefore, for each $u \in (\ker A)^\perp$ there exists $v \in \ell^2(\mathbb{N})$ satisfying (69). Moreover, we calculate by (66) that

$$\begin{aligned}(Bv, v)_{\ell^2(\mathbb{N})} &= (Bv, P_{B,\perp}v)_{\ell^2(\mathbb{N})} = (Bv, \tilde{B}^{-1}Au)_{\ell^2(\mathbb{N})} = (Au, \tilde{B}^{-1}Au)_{\ell^2(\mathbb{N})}\\ &= (Au, P_{A,\perp}\tilde{B}^{-1}Au)_{\ell^2(\mathbb{N})} = (Au, \tilde{A}^{-1}Au)_{\ell^2(\mathbb{N})} = (Au, u)_{\ell^2(\mathbb{N})}.\end{aligned}$$

Thus (68) holds true. We hence prove that $\mathcal{H}_{K_A} \preceq \mathcal{H}_{K_B}$. ■

To close this section, as an application of Theorem 32, we consider kernels of finite dimensional feature spaces. Let $n \le m$ be two positive integers and A, B hermitian and strictly positive definite matrices of sizes $n \times n$ and $m \times m$, respectively. Suppose that ϕ_j, $j \in \mathbb{N}_m$ form a sequence of linearly independent functions on X. The kernels we consider are

$$G_A(x,y) := \sum_{j,k \in \mathbb{N}_n} A_{jk}\phi_k(x)\overline{\phi_j(y)},\ G_B(x,y) := \sum_{j,k \in \mathbb{N}_m} B_{jk}\phi_k(x)\overline{\phi_j(y)},\ \ x, y \in X. \tag{70}$$

We remark that a kernel has a finite dimensional feature space if and only if its RKHS has finite dimension. It was proven in Aronszajn (1950) that a RKHS is finite dimensional if and only if its reproducing kernel has the form of (70). The following corollary is a direct consequence of Theorem 32.

Corollary 33 *Let kernels G_A, G_B be defined by (70). Then $\mathcal{H}_{G_A} \preceq \mathcal{H}_{G_B}$ if and only if B^{-1} is an augmentation of A^{-1}, namely, $B^{-1}_{jk} = A^{-1}_{jk}$, $j, k \in \mathbb{N}_n$. In particular, if G_A, G_B have the form*

$$G_A(x,y) := \sum_{j \in \mathbb{N}_n} a_j\phi_j(x)\overline{\phi_j(y)},\ G_B(x,y) := \sum_{k \in \mathbb{N}_m} b_k\phi_k(x)\overline{\phi_k(y)},\ \ x, y \in X$$

for some positive constants a_j, b_k then $\mathcal{H}_{G_A} \preceq \mathcal{H}_{G_B}$ if and only if $a_j = b_j$ for each $j \in \mathbb{N}_n$. In both cases, if $\mathcal{H}_{G_A} \preceq \mathcal{H}_{G_B}$ then G_B is a nontrivial refinement kernel for G_A if and only if $m > n$.

Acknowledgments

Yuesheng Xu was supported in part by the US National Science Foundation under grants CCR-0407476 and DMS-0712827, by the Natural Science Foundation of China under grants 10371122 and 10631080, by the Education Ministry of the People's Republic of China under the Changjiang Scholar Chair Professorship Program through Sun Yat-sen University, and by Institute of Mathematics, Academy of Mathematics and System Science, Chinese Academy of Sciences.

References

N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68:337–404, 1950.

S. Bochner. *Lectures on Fourier Integrals with an Author's Supplement on Monotonic Functions, Stieltjes Integrals, and Harmonic Analysis*. Annals of Mathematics Studies 42, Princeton University Press, New Jersey, 1959.

O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

Q. Chen, C. A. Micchelli, S. Peng and Y. Xu. Multivariate filters banks having a matrix factorization. *SIAM J. Matrix Anal. Appl.*, 25:517-531, 2003.

Q. Chen, C. A. Micchelli and Y. Xu. On the matrix completion problem for multivariate filter bank construction. *Adv. Comput. Math.*, 26:173–204, 2007.

J. B. Conway. *A Course in Functional Analysis*. 2nd Edition, Springer-Verlag, New York, 1990.

F. Cucker and S. Smale. On the mathematical foundations of learning. *Bull. Amer. Math. Soc.*, 39:1–49, 2002.

I. Daubechies. *Ten Lectures on Wavelets*. CBMS-NSF Regional Conference Series in Applied Mathematics 61, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1992.

T. Evgeniou, M. Pontil and T. Poggio. Regularization networks and support vector machines. *Adv. Comput. Math.*, 13:1–50, 2000.

C. H. FitzGerald, C. A. Micchelli and A. Pinkus. Functions that preserve families of positive semidefinite matrices. *Linear Algebra Appl.*, 221:83–102, 1995.

C. Gasquet and P. Witomski. *Fourier Analysis and Applications*. Springer-Verlag, New York, 1999.

G. H. Golub and C. F. van Loan. *Matrix Computations*. 3rd Edition, Johns Hopkins University Press, Baltimore, MD, 1996.

T. N. T. Goodman and S. L. Lee. Wavelets of multiplicity r. *Trans. Amer. Math. Soc.*, 342:307–324, 1994.

L. Grafakos. *Classical and Modern Fourier Analysis*. Prentice Hall, New Jersey, 2004.

H. Hochstadt. *Integral Equations*. Wiley, New York, 1973.

R. Q. Jia. Characterization of smoothness of multivariate refinable functions in Sobolev spaces. *Trans. Amer. Math. Soc.*, 351:4089–4112, 1999.

R. Q. Jia, S. D. Riemenschneider and D. X. Zhou. Smoothness of multiple refinable functions and multiple wavelets. *SIAM J. Matrix Anal. Appl.*, 21:1–28, 1999.

G. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *J. Math. Anal. Appl.*, 33:82–95, 1971.

J. Mercer. Functions of positive and negative type and their connection with the theory of integral equations. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.*, 209: 415–446, 1909.

C. A. Micchelli and M. Pontil. Learning the kernel function via regularization. *Journal of Machine Learning Research*, 6: 1099–1125, 2005.

C. A. Micchelli and M. Pontil. On learning vector-valued functions. *Neural Comput.*, 17:177–204, 2005.

C. A. Micchelli and T. Sauer. Regularity of multiwavelets. *Adv. Comput. Math.*, 7:455–545, 1997.

C. A. Micchelli and Y. Xu. Using the matrix refinement equation for the construction of wavelets on invariant sets. *Appl. Comput. Harmon. Anal.*, 1:391–401, 1994.

C. A. Micchelli, Y. Xu and P. Ye. Cucker Smale learning theory in Besov spaces. In *Advances in Learning Theory: Methods, Models and Applications*, pages 47–68, IOS Press, Amsterdam, The Netherlands, 2003.

C. A. Micchelli, Y. Xu and H. Zhang. Universal kernels. *Journal of Machine Learning Research*, 7:2651–2667, 2006.

S. Mukherjee, P. Niyogi, T. Poggio and R. Rifkin. Learning theory: Stability is sufficient for generalization and necessary and sufficient for empirical risk minimization. *Adv. Comput. Math.*, 25:161–193, 2006.

J. R. Munkres. *Topology*. 2nd Edition, Prentice Hall, Upper Saddle River, New Jersey, 2000.

R. Opfer. Multiscale kernels. *Adv. Comput. Math.*, 25: 357–380, 2006.

A. Rakotomamonjy and S. Canu. Frames, reproducing kernels, regularization and learning. *Journal of Machine Learning Research*, 6:1485–1515, 2005.

T. J. Rivlin. *Chebyshev Polynomials*. 2nd Edition, Wiley, New York, 1990.

W. Rudin. *Real and Complex Analysis*. 3rd Edition, McGraw-Hill, New York, 1987.

S. Saitoh. *Theory of Reproducing Kernels and its Applications*. Pitman Research Note in Mathematics Series 189, Longman, UK, 1988.

I. J. Schoenberg. Metric spaces and completely monotone functions. *Ann. of Math. (2)*, 39:811–841, 1938.

I. J. Schoenberg. Positive definite functions on spheres. *Duke. Math. J.*, 9: 96–108, 1942.

B. Schölkopf, R. Herbrich and A. J. Smola. A generalized representer theorem. In *Proceeding of the 14th Annual Conference on Computational Learning Theory and the 5th European Conference on Computational Learning Theory*, pages 416–426, Springer-Verlag, London, UK, 2001.

B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, Mass, 2002.

J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge, 2004.

S. Smale and D. X. Zhou. Estimating the approximation error in learning theory. *Anal. Appl.*, 1:17–41, 2003.

E. M. Stein and G. Weiss. *Introduction to Fourier Analysis on Euclidean Spaces*. Princeton University Press, New Jersey, 1971.

I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, 2001.

I. Steinwart and C. Scovel. Fast rates for support vector machines using Gaussian kernels. In *Proceeding of the 18th Annual Conference on Learning Theory* (COLT 05), pages 279–294, Bertinoro, 2005.

H. Sun. Mercer theorem for RKHS on noncompact sets. *J. Complexity*, 21:337–349, 2005.

V. N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

G. Wahba. Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV. In *Advances in Kernel Methods–Support Vector Learning*, pages 69–86, MIT Press, Cambridge, Mass, 1999.

C. Walder, B. Schölkopf and O. Chapelle. Implicit surface modelling with a globally regularised basis of compact support. *Computer Graphics Forum*, 25:635–644, 2006.

Y. Wang. Wavelets, tiling, and spectral sets. *Duke Math. J.*, 114:43–57, 2002.

D. V. Widder. *The Laplace Transform*. Princeton University Press, Princeton, 1941.

Y. Xu and H. Zhang. Refinable kernels. *Journal of Machine Learning Research*, 8:2083–2120, 2007.

Y. Ying and D. X. Zhou. Learnability of Gaussians with flexible variances. *Journal of Machine Learning Research*, 8:249–276, 2007.

T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Ann. Statis.*, 32:56–85, 2004.